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.

May 6, 2024

CONTENTS

LIST OF FIGURES

LIST OF TABLES

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 consumer electronics market is a rapidly evolving and highly competitive sector, characterised by constant technological advancements, shifting consumer preferences, and dynamic pricing strategies. As such, businesses and researchers require comprehensive and up-to-date data on product offerings and market trends to navigate this complex landscape effectively. However, the lack of publicly available datasets that provide detailed information on consumer electronics products sold on major e-commerce platforms poses a significant challenge for those seeking insights into this industry.

To fill the gap, this research project used web scraping to create a dataset analysing AliExpress market trends and consumer preferences in the *"Consumer Electronics"* product category of the AliExpress online retail platform. Specifically, AliExpress was selected as the primary data source for this project due to several key factors that align with this project's research goals and the unique characteristics of the consumer electronics market."

- **Product Variety:** AliExpress offers an extensive range of consumer electronics products, from smartphones and laptops to smart home devices and wearables. This diverse product catalogue allows for a comprehensive analysis of the consumer electronics market, capturing the breadth of available products and enabling the identification of emerging trends across various subcategories.

- **Seller Diversity:** The AliExpress platform hosts sellers from different geographic regions, including manufacturers, wholesalers, and small businesses. This diversity of sellers provides a more representative sample of the global consumer electronics market, allowing for the examination of pricing strategies, product quality, and market competitiveness across different seller types and regions.

- **Data Accessibility:** Although AliExpress does not provide a public API, its website structure allows straightforward web scraping. Product information, such as titles, prices, sales volumes, and store details, is generally well-structured and consistent

across product pages, enabling the efficient extraction of relevant data. Moreover, AliExpress's terms of service do not explicitly prohibit web scraping for research purposes, making it a more accessible data source than other platforms with stricter restrictions.

➢ **Market Relevance:** With its global reach, vast user base, and significant market share, AliExpress plays a pivotal role in shaping the consumer electronics industry. The platform's sales volume and customer feedback provide valuable insights into consumer demand, preferences, and satisfaction levels, making it a highly relevant data source for understanding the dynamics of the consumer electronics market.

To further justify the choice of AliExpress as the primary data source, this research project conducted a comparative analysis of several major e-commerce platforms, including Temu, Amazon, and eBay. The study revealed that while these platforms offer a wide range of consumer electronics products, AliExpress stands out in its product variety, seller diversity, and data accessibility. For instance, Temu, a relatively new player in the e-commerce space, has a more limited product catalogue and seller base than AliExpress. While offering a vast array of products, Amazon and eBay have more stringent web scraping restrictions and complex website structures, making data extraction more challenging and potentially less comprehensive.

In contrast, AliExpress provides a rich and accessible data source well-suited for studying the consumer electronics market. The platform's diverse product offerings, global seller network, and relatively structured website layout enable the creation of a comprehensive dataset that captures the nuances and dynamics of this rapidly evolving industry.

Table 1.1
Comparative Analysis of E-Commerce Platforms

Feature	**AliExpress**	**Temu**	**Amazon**	**eBay**
Product Variety	High	Moderate	High	High
Seller Diversity	High	Low	Moderate	High
Data Accessibility	High	Low	Moderate	Moderate
Market Relevance	High	Low	High	High
Legal Restrictions	Low	High	Moderate	Moderate

After careful consideration of various retail platforms, the e-commerce retail platform AliExpress was chosen due to its superior features compared to the other three e-commerce platforms considered (i.e., Temu, Amazon, eBay) and its concrete and strong alignment with the requirements, objectives, and scope of this web scraping project. Moreover, Table 1.1 presented above highlights the reasons underlying this decision.

Furthermore, by leveraging web scraping techniques to extract data from AliExpress, this research project aimed to create a unique and valuable resource for researchers, businesses, and market analysts. The dataset encompasses various aspects of products from the *"Consumer Electronics"* product category of the AliExpress website, including titles, prices, sales volumes, and store details, enabling a multifaceted analysis of product offerings, market trends, and pricing strategies.

The primary objective was to develop a robust web scraping methodology to efficiently extract relevant data from AliExpress product listings and transform it into a structured dataset suitable for further analysis. This dataset empowers its future users to:

- ➢ Identify popular product categories and emerging technologies in the consumer electronics space, facilitating trend forecasting and product development strategies.
- ➢ Analyse pricing strategies employed by different sellers and brands, providing insights into market competitiveness and potential opportunities for price optimisation.
- ➢ Compare product offerings and prices across different regions or countries, facilitating market entry decisions and global expansion strategies.

➢ Develop machine learning models for product recommendation or demand forecasting, leveraging the rich data available in the dataset.

The AliExpress consumer electronics dataset aims to fill a critical gap in the availability of comprehensive and accessible data on this dynamic industry. By providing a valuable resource for researchers and businesses alike, this dataset has the potential to drive innovation, inform strategic decision-making, and contribute to a deeper understanding of the global consumer electronics market. The dataset's unique focus on AliExpress, with its diverse product offerings and global reach, ensures that the insights derived from this resource will be comprehensive and relevant to the industry's current state.

In conclusion, the primary motivation behind creating this dataset was to empower researchers, businesses, and market analysts with a comprehensive and accessible resource for studying the consumer electronics market on AliExpress. By leveraging web scraping techniques to extract valuable data from one of the world's largest and most diverse e-commerce platforms, this dataset aims to provide a unique and powerful tool for gaining insights into market trends, consumer preferences, and competitive dynamics in the rapidly evolving consumer electronics industry.

Chapter 2

COMPOSITION

What do the instances that comprise the dataset represent?

The instances comprising the AliExpress webs scraped consumer electronics product dataset represent individual consumer electronics products listed on the AliExpress website. Each instance corresponds to a unique product, and the attributes associated with each instance provide various details about the product, such as its title, price, store information, and more.

How many instances are there in total?

The dataset contains 900 instances, representing 900 unique products scraped from the AliExpress website across 18 pages of product listings from the *"Consumer Electronics"* product category.

Extraction Design

To enhance the robustness and comprehensiveness of the data extraction process, the Python code script was designed to navigate through multiple pages of results from the *"Consumer Electronics"* product category on the AliExpress website, ensuring a diverse array of product listings across various price ranges, sales figures, and product subcategories within the broader *"Consumer Electronics"* product category on the AliExpress website.

The Python code script developed for this web scraping research project incorporates a sequential pagination logic, starting from the first search results page and iteratively moving to subsequent pages until the desired sample size is reached or no more pages are available. The Python code script extracts relevant product information for each *"Consumer Electronics"* product category web page, including the product URL, title, price, currency, units sold, thumbnail image URL, and AliExpress store details (i.e., URL, name, and other identifiers).

To handle the dynamic loading of content on the AliExpress results pages, the script employs Selenium's WebDriver to simulate user interactions, including scrolling and click-

ing on elements. This enables the Python code script to access and extract data from sections of the page that may not be immediately visible or loaded upon initial page load.

During the data extraction process, several challenges were encountered and addressed:

- **IP Blocking:** To mitigate the risk of IP blocking due to excessive requests, the Python code script incorporated random delays between requests using the `time.sleep()` function. The delays were set to random intervals within a reasonable range to simulate human browsing behaviour and avoid triggering AliExpress's anti-scraping mechanisms.
- **AJAX-loaded Content:** Some product information on AliExpress is loaded dynamically through AJAX requests. To handle this, the script used Selenium's `WebDriverWait` to wait for specific elements on the page before extracting the product data. This ensured the Python code script waited for the necessary content to load before proceeding with the extraction.
- **Pagination:** AliExpress employs a complex pagination system, with dynamically loaded *"Load More"* buttons and AJAX-based pagination. The script handled this by monitoring the presence of the *"Load More"* button and clicking it when available to load additional products. It also kept track of the current page number and updated the URL accordingly so that the web scraping algorithm could navigate through the results pages on the AliExpress website.
- **Respect for Website Terms of Service:** To ensure compliance with AliExpress's terms of service and maintain a respectful crawl rate, the script implemented configurable delay parameters to control the time between requests. The delays were set to reasonable values to avoid overloading AliExpress's servers and to mimic human browsing patterns. Additionally, the script adhered to the directives specified in AliExpress's `robots.txt` file, ensuring it only accessed pages allowed for web scraping.

Research Validity and Balance

In web scraping for marketing insights, particularly in extracting consumer electronics data from AliExpress, maintaining a balance between research validity, technical feasibility, and

legal/ethical considerations is paramount. This project has been meticulously designed to optimise data collection processes, ensuring robustness and integrity while adhering to ethical and legal standards.

Optimising Research Validity: The methodology for this web scraping research project aimed to capture a comprehensive and representative sample of the consumer electronics market. By targeting a diverse range of products (i.e., 500-1000 product listings), the project minimised selection bias and enhanced the generalisability of the findings. The extraction logic was consistently applied across different product pages to ensure uniformity in data collection, thereby maintaining high research validity.

Addressing Technical Challenges: The Python script developed for this AliExpress web scraping research project facilitated effective data extraction from the AliExpress online retail platform by efficiently navigating multiple pages on the AliExpress website. This script was equipped to handle various technical challenges, such as dynamic content loading and pagination complexities. Techniques like Selenium's WebDriver were utilised to simulate user interactions, ensuring that all relevant product information was accurately captured. This approach not only streamlined the data collection process but also safeguarded against common pitfalls like IP blocking and AJAX-loaded content, which could compromise data integrity.

Legal and Ethical Compliance: This AliExpress web scraping research project strictly adhered to AliExpress's terms of service and the directives outlined in their `robot.txt` file. The Python code script accessed only publicly available information and was programmed to avoid bypassing website restrictions. To respect user privacy and comply with ethical standards, measures were implemented to anonymise any incidentally collected personal data, ensuring that no identifiable information about individuals was stored. Furthermore, the dataset was used exclusively for research purposes, with stringent safeguards to prevent misuse.

Mitigating Ethical Risks: The Python code script developed for this AliExpress product data web scraping project incorporated delay timers between requests to mimic human browsing patterns and prevent overloading the AliExpress servers. These measures were crucial in maintaining a respectful crawl rate and avoiding disrupting the website's normal

operations. Additionally, user agent headers identified the nature of the data collection as part of a research project, promoting transparency and accountability.

By precisely designing the data collection framework and employing a strategic approach to web scraping, this research project successfully navigates the intricate balance between maximising research validity, overcoming technical hurdles, and adhering to legal and ethical guidelines. The resulting AliExpress web-scraped consumer electronics product data dataset is a valuable asset for analysing market trends and consumer preferences and is a testament to the rigorous standards upheld throughout the research process.

Does the dataset contain all possible instances or is it a sample of instances from a larger set?
The dataset is a sample of instances from the larger set of all consumer electronics products available on the AliExpress website. The Python script web scraped the data collected from 18 product listing pages, yielding 900 instances. However, this represents only a subset of the total number of product listings available in the *"Consumer Electronics"* product category on the AliExpress online retail platform.

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

- `id`: The unique product ID assigned by AliExpress.
- `url`: The URL of the product page on the AliExpress website.
- `type`: The product type (e.g., `natural` or `ad`).
- `title`: The title or name of the product.
- `price`: The price of the product.
- `currency`: The currency in which the price is listed.
- `trade`: Trade information, such as the number of units sold.
- `thumbnail`: The URL of the product thumbnail image.
- `store_url`: The URL of the AliExpress store selling the product.

- `store_name`: The name of the store selling the product.
- `store_id`: The unique ID of the store selling the product.
- `store_ali_id`: The AliExpress member ID of the store selling the product.

Is there a label or target associated with each instance?
No explicit labels or targets are associated with each instance in the dataset. The dataset is primarily intended for exploratory analysis and market trend identification rather than for tasks such as classification or prediction that would require labelled data.

Is any information missing from individual instances?
Based on the provided Python script and the resulting dataset, some attributes, such as the `trade` attribute, may be missing for certain instances. The script uses the `get()` method with a default value when extracting the `trade` information, indicating that this attribute may not be available for all products.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?
No, the dataset does not contain any explicit relationships between individual instances. Each instance represents a standalone consumer electronics product, and the attributes provided do not establish any direct connections or relationships between the products.

Are there recommended data splits (e.g., training, development/validation, testing)?
As the dataset is primarily intended for exploratory analysis and market trend identification, there are no recommended data splits for tasks such as training, validation, and testing. The entire dataset can be used to analyse patterns, trends, and insights related to products in the *"Consumer Electronics"* product category on the AliExpress website.

Did you collect the data from the source(s)?
Yes, the data were collected directly from the AliExpress website using a custom Python script. The script utilised the `Selenium` and `Parsel` libraries to automate the web scraping process, navigating through the search results pages and extracting relevant product

information.

Chapter 3

DATA INSPECTION

The data inspection process, a pivotal step in guaranteeing the dataset's quality and reliability for accurate analysis and meaningful insights, is thoroughly detailed in this section. The primary goals of data inspection are to comprehend the data distribution, detect any anomalies or outliers, and evaluate the overall dataset quality. This section provides a comprehensive overview of the procedures executed in the *R* code script to inspect the AliExpress consumer electronics dataset, encompassing data loading, cleaning, and analysis.

Data Loading and Cleaning

The dataset was loaded into the *R* environment using the `read_csv()` function from the `readr` package. The `head()` and `str()` functions were then used to examine the basic structure and content of the dataset, providing an overview of the variables and their data types.

To address missing values, the `is.na()` function was applied to each column, revealing that the `trade` variable contained missing data. These missing values were subsequently replaced with zeros using the following code:

```
product_data$trade_numeric[is.na(product_data$trade_numeric)] <- 0
```

Furthermore, the `price` and `trade` variables were converted from character to numeric data types using the `gsub()` and `as.numeric()` functions, enabling mathematical operations and statistical analysis:

```
product_data$price_numeric <- as.numeric(gsub("[^0-9\\.]", "",
    product_data$price))
product_data$trade_numeric <- as.numeric(gsub("[^0-9\\.]", "",
    product_data$trade))
```

Statistical Analysis

Various statistical methods were employed to gain a deeper understanding of the dataset. Summary statistics, including mean, median, mode, standard deviation, and range, were calculated for each variable using the `summary()` function. This provided insights into the central tendencies and dispersion of the data. The summary statistics for the `price_usd` variable are presented in Table 3.1 below.

Table 3.1
Summary Statistics Table for Price Range Analysis

Property	**Min.**	**1st Qu.**	**Median**	**Mean**	**3rd Qu.**	**Max.**
Value	0.71	0.99	2.36	7.58	8.01	51.94

The statistics presented in Table 3.1 above indicate a wide range of prices for consumer electronics products on AliExpress, with a median price of $2.36 and a mean price of $7.58, suggesting the presence of some higher-priced items skewing the average.

Data Quality Assessment

A thorough data quality assessment was conducted to ensure the reliability of the dataset. This involved identifying and handling missing values, outliers, and duplicate entries. As mentioned earlier, missing values in the `trade` variable were replaced with zeros to facilitate analysis.

Outliers were detected using the interquartile range (IQR) method, identifying values below `Q1 - 1.5` $\times$ `IQR` or above `Q3 + 1.5` $\times$ `IQR`. For the `price_usd` variable, the following code was used to identify outliers:

```
Q1 <- quantile(product_data$price_usd, 0.25)
Q3 <- quantile(product_data$price_usd, 0.75)
IQR <- Q3 - Q1
outliers <- product_data$price_usd < (Q1 - 1.5 * IQR) | product_data$price_usd >
    (Q3 + 1.5 * IQR)
```

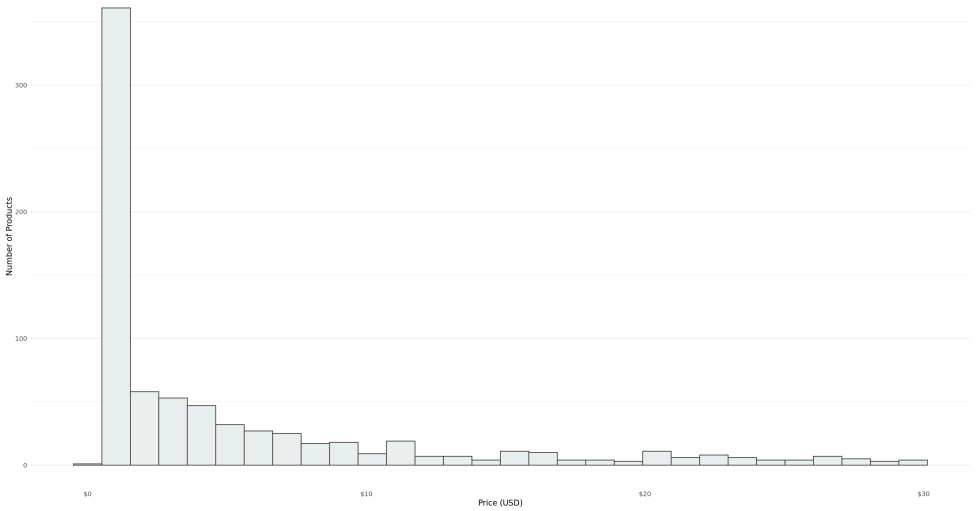
Duplicate entries were identified using the `duplicated()` function, and no duplicates were found in the dataset.

Data Visualisation

Various visualisations were created using the `ggplot2` library for *R* to explore the dataset further and identify patterns or trends. These visualisations included histograms, box plots, and scatter plots, which helped illustrate the range and spread of key variables. For instance, the price distribution plot that is presented in Figure 3.1 reveals a right-skewed distribution, with the majority of consumer electronics products priced below $30:

```
# Analyse the price distribution to identify pricing strategies and ranges.
price_distribution_plot <- ggplot(translated_product_data, aes(x = price_usd)) +
  geom_histogram(bins = 30, fill = "#E9EEEE", color = "#383535") +
  scale_x_continuous(labels = scales::dollar_format()) +
  labs(title = "Distribution of Prices in USD", x = "Price (USD)", y = "Number
    of Products") +
  theme_minimalist()
price_distribution_plot
```

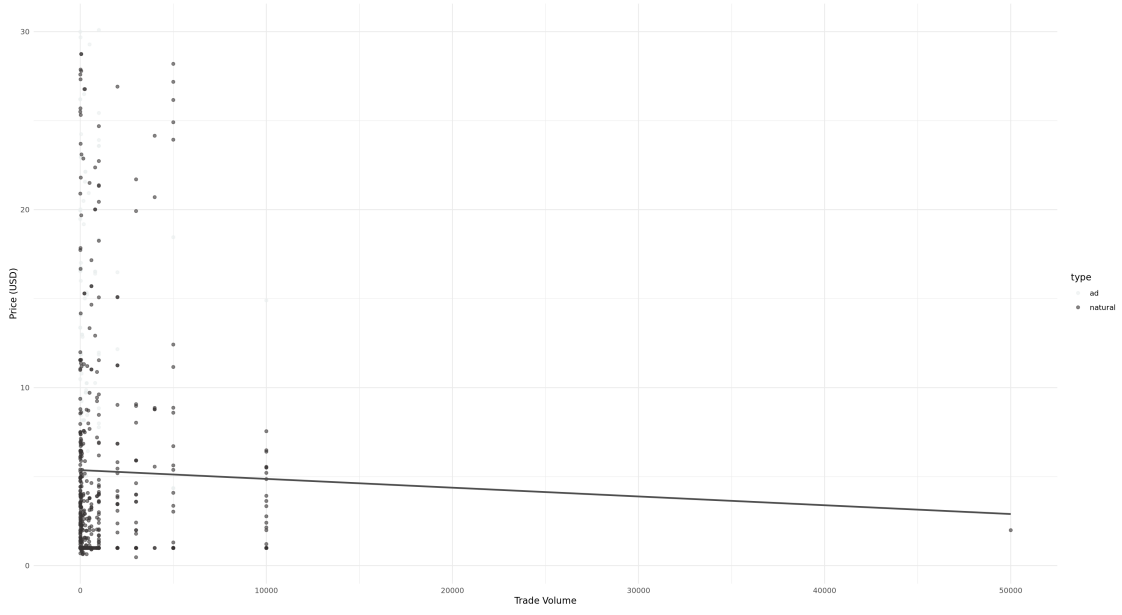
Figure 3.1
Distribution Plot of Prices Expressed in USD



The scatter plot of trade volume versus price in USD that is presented in Figure 3.2 below shows a weak negative correlation, suggesting that lower-priced consumer electronics products sold on the AliExpress online retail platform tend to have slightly higher sales volumes:

```
# Explore the relationship between `trade_numeric` and `price_usd` to see if
    higher trade correlates with pricing.
trade_price_correlation <- ggplot(translated_product_data, aes(x =
    trade_numeric, y = price_usd)) +
  geom_point(aes(color = type), alpha = 0.6) +
  geom_smooth(method = "lm", se = FALSE, color = "gray30") +
  labs(title = "The Correlation Between Trade Volume and Price", x = "Trade
    Volume", y = "Price (USD)") +
  theme_minimal() +
  scale_color_manual(values = c("#E9EEEE", "#383535", "#73787C"))
trade_price_correlation
```

Figure 3.2
Correlation Plot Between Trade Volume and Price Expressed in USD



Bias Assessment

However, it is essential to acknowledge that the dataset only represents a snapshot of products available on AliExpress during scraping and may not capture the full breadth of the *"Consumer Electronics"* product category on the AliExpress website. The dataset could be combined with other e-commerce platforms or market research sources to mitigate this limitation and provide a more comprehensive market view.

Conclusion

The data inspection process revealed valuable insights into the AliExpress consumer electronics dataset. The dataset was complete, with only a few missing values in the `trade` variable. The wide range of prices and outliers highlighted the diversity of products available on the platform.

The visualisations clarified the data distribution and the relationships between key variables. The histogram of prices showed a right-skewed distribution, while the scatter plot indicated a weak positive correlation between trade volume and price.

However, the limitations of the dataset, such as its snapshot nature and the focus on a single e-commerce platform, should be considered when drawing conclusions or making generalisations.

Overall, the thorough data inspection process employed in this study enhances the robustness of the research findings and provides confidence in the dataset's ability to support meaningful insights into the consumer electronics market on AliExpress. The insights gained from this process will guide the subsequent analysis and inform the overall research strategy of the online data collection and management project.

Chapter 4

DATA COLLECTION PROCESS

How were the data associated with each instance acquired?
The data associated with each instance in the AliExpress consumer electronics product dataset were directly observable, as the data were web-scraped from the AliExpress website using a custom Python code script. The code script extracted raw text data for each product, including the product title, price, number of units sold, store information, and other relevant details. No data were reported by subjects or indirectly inferred/derived from other sources, ensuring the dataset's accuracy and reliability.

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?
A custom Python script was developed to collect the data from the AliExpress website. The script utilised the `Selenium` library for web scraping and automation, the `Parsel` library for parsing HTML content and extracting relevant data using XPath and regular expressions, and the `HTTPX` library for making `HTTP` requests to navigate the AliExpress website.

The data extraction process faced several technical challenges due to the complex site architecture of AliExpress, which heavily relies on `AJAX` calls and lazy-loaded content. To overcome these challenges, the code script was designed to handle dynamic elements and ensure all relevant product data was captured. Pagination was managed by constructing the `URL` for each consumer electronics product category web page and iterating through the desired number of pages. Below, I have included a simplified representation of pagination handling in the Python code script that was developed and, in turn, used to web scrape the consumer electronics product data from the AliExpress website.

```
# Define the total number of pages.
    num_pages = 18
    # Initialise an empty list, named ``all_product_data``, to store the product
    data.
```

```
all_product_data = []
# Loop through each page.
for page_num in range(1, num_pages + 1):
    url =
f"https://www.aliexpress.com/category/44/consumer-electronics.html?page=
    {page_num}"
    resp = httpx.get(url, follow_redirects=True)
    product_data = parse_search(resp)
    all_product_data.extend(product_data)
```

The Python code script relied on the presence of a specific `<script>` tag containing the product data in `JSON` format, which was extracted using regular expressions and parsed to obtain the relevant product details.

To ensure data quality and completeness, the final version of the code script implemented various checks and error-handling mechanisms. Key elements like prices and titles were verified before recording a product, and failed requests were retried several times before logging an error and moving on. The Python code script was carefully designed, tested, and continuously monitored to ensure accurate and consistent data extraction.

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

The AliExpress consumer electronics dataset is a sample from the larger set of all consumer electronics products available on the AliExpress platform. A sequential sampling strategy was employed, where the code script collected product data from the first 18 pages of search results in the *"Consumer Electronics"* category on AliExpress. This approach aimed to capture a representative sample of the category's most relevant and popular products.

While a fully randomised sampling strategy was initially considered, the sequential method was chosen due to limited access to AliExpress's underlying product database. By focusing on the top search results across multiple pages of the *"Consumer Electronics"* product category on the AliExpress website, the resulting data sample is expected to provide valuable insights into the consumer electronics market on the AliExpress online retail platform for other users in the future. The code script can be easily modified to adjust the number of pages web scraped or target different product categories on the AliExpress online retail website, allowing flexibility in the sampling approach based on research requirements.

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)?
The data collection process was conducted by the author of this online data collection and management research project report without the involvement of external parties such as crowd workers or contractors. As a result, no compensation was provided for the data collection efforts.

Over what time frame were the data collected? Does this time frame match the creation time frame of the data associated with the instances (e.g., the recent crawl of old news articles)? If not, please describe the time frame in which the data associated with the instances were created.
The data were collected in one day in May of 2024, which aligns with the creation time frame of the data associated with the instances. The dataset represents a snapshot of product data from the *"Consumer Electronics"* product category, which was available on the AliExpress website during this specific period of a single day, ensuring the relevance and timeliness of the collected data for analysing current market trends and consumer preferences.

Were any ethical review processes conducted (e.g., by an institutional review board)? If so, please describe these review processes, including the outcomes, and provide a link or other access point to any supporting documentation.
The data collection process did not involve human subjects or personally identifiable information and, therefore, did not require a formal ethical review process by an institutional review board. However, this web scraping project adhered to fundamental ethical principles throughout the study, ensuring that the data collection and analysis were conducted responsibly and transparently without causing any harm or infringing upon the rights of any individuals or entities involved.

Does the dataset relate to people? If not, you may skip the remainder of the questions in this section.
The AliExpress consumer electronics dataset does not directly relate to people, as it focuses on product information and does not contain any personally identifiable information or sensitive data about individuals. As such, the remaining questions in this section, which

pertain to data collection from individuals, do not apply to this dataset.

Chapter 5

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 to enhance its quality and prepare it for analysis:

- **Title Translation:** The product titles in the `title` data column of the `product_data` data frame were translated from their original languages to English using the `googleLanguageR` library. This step aimed to enhance the dataset's usability and facilitate analysis by ensuring all product titles were in a common language. The translation process involved setting the Google Cloud Translation API key and defining a function to translate the titles using the `gl_translate()` function. This function was then applied to the `product_data` data frame, resulting in a new data frame, `translated_product_data`, with the translated titles stored in the `translated_title` column.

- **Missing Value:** The initial assessment revealed 36 instances with missing values in the `trade` data column, representing the number of units sold. These missing values were imputed with the value 0, assuming a missing value indicates no sales recorded at the time of web scraping.

- **Data Cleaning:** The `trade` column contained values like `1,000+ sold` and `50,000+ sold`. To facilitate numerical analysis, these values were cleaned by removing the `+` symbol and the text `sold` using regular expressions. The resulting values were then converted to numeric format.

- **Currency Standardisation:** While the majority of products were listed in USD, a minority of the products were listed in Philippine peso (PHP). To ensure consistency for price analysis, the `price_usd` data column was created. Prices expressed in PHP

were converted to USD using a fixed conversion rate of 50 (i.e., the approximate rate at the time of data collection).

➢ **Thumbnail URL Creation:** The `thumbnail` data column contained only the image file path. To create complete URLs for accessing product thumbnail images, a new column named `thumbnail_url` was created by concatenating the base URL of AliExpress pictures with the file path.

The specific *R* code used for these preprocessing steps can be found in the `oDCM_Resit_WebScraping_AliExpress_Consumer_Electronics_Product_Data_R_Code_Script_Preprocessing_Cleaning_Labelling.R` file provided with the 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.

Yes, the raw data extracted directly from the AliExpress website were preserved in the `Raw_AliExpress_Consumer_Electronics_Data.csv` CSV data file. This file contains the original data with 900 instances, including the `trade` data column with its original format and the `price` data column with the original currencies.

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

The *R* code script used for preprocessing, cleaning, and labelling is available in the R source code file as referenced above. This R code script demonstrates the exact steps to transform the raw data into the translated and, in turn, cleaned dataset.

Any other comments?

The preprocessing steps were crucial in improving the quality and usability of the AliExpress consumer electronics dataset. These steps addressed missing values, cleaned and standardised the format of sales figures, ensured consistent currency for price analysis, and enriched the data with complete thumbnail URLs. As a result, the dataset is now well-suited for various analytical tasks, including exploring product categories, price ranges, seller details, and potential correlations between variables.

The decision to impute missing values in the `trade` column with 0 was based on the assumption that missing values indicate no recorded sales at the time of web scraping. This assumption may require further investigation and potentially alternative imputation methods depending on the specific research questions and the availability of additional information.

The fixed conversion rate of 50 for PHP to USD was chosen based on the approximate exchange rate at the time of data collection. For more accurate analysis, real-time or historical exchange rates should be used to convert prices to a common currency.

Overall, the preprocessing steps significantly enhanced the quality and analytical value of the AliExpress consumer electronics dataset, paving the way for reliable and insightful exploration of market trends and consumer preferences in the online electronics market.

Chapter 6

USES

Has the dataset been used for any tasks already? If so, please provide a description.

This dataset has not been used for specific tasks yet since it was collected as part of an online data collection and management research project on web scraping and online data collection, and its primary purpose has been for educational and skill development purposes.

Is there a repository linking to any papers or systems that use the dataset? If so, please provide a link or other access point.

As the AliExpress consumer electronics dataset is newly created and has not been published, there is no repository linking to papers or systems that use this dataset.

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

The AliExpress consumer electronics dataset offers a wealth of potential applications for market research, trend analysis, and business strategy:

- **Identifying Trending Products and Emerging Technologies:** By analysing product titles and descriptions, researchers, analysts, and business decision-makers can uncover popular features, specifications, and selling points, gauging their prevalence in the current market.
- **Investigating Pricing Strategies:** The dataset enables pricing comparisons across various market segments, revealing insights into different sellers' competitive dynamics and pricing patterns.
- **Studying Customer Preferences:** Combining sales volume data with pricing information allows researchers, analysts, and business decision-makers to examine the relationship between price and demand, inferring features that resonate most with customers.

- **Comparing Offerings Across Regions:** The currency information can explore how consumer electronics trends and preferences vary by country or region, informing market expansion strategies.
- **Developing Machine Learning Models:** The rich product attributes and sales data are a foundation for training recommendation systems, demand forecasting tools, and other artificial intelligence (AI) applications.

Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labelled that might impact future uses? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?

There are a few aspects of the dataset's composition and preprocessing that users should be mindful of:

- **Missing Sales Volume Data:** A significant portion of the *trade* column, indicating sales volume, contains missing values. Analyses or models heavily reliant on this variable may be biased or inaccurate. Users should assess the extent of missing data and consider appropriate imputation techniques or alternative proxy variables.
- **Temporal Limitations:** The dataset represents a snapshot of product listings at a specific point in time and may not capture long-term trends or seasonal fluctuations. Combining this data with other sources or conducting regular updates could help mitigate this limitation.
- **Platform Specificity:** As the dataset focuses on a single e-commerce platform, its generalisability to the broader consumer electronics market may be limited. Users should be cautious when drawing conclusions or making business decisions based solely on this dataset.
- **Preprocessing Transformations:** The conversion of prices to numeric values and filling missing trade volumes with zeros could impact specific analyses. Users should review these transformations and assess their alignment with particular research objectives, considering alternative techniques when appropriate.

To mitigate potential risks or harms, dataset consumers should handle the data responsibly, combining it with other relevant sources and domain expertise to validate critical findings before making strategic decisions. Regularly updating the dataset and expanding its coverage to include additional platforms could enhance its representativeness and utility.

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

Given the focus on product-level information and the absence of personally identifiable data, there are no major ethical concerns surrounding the use of this dataset. However, users should refrain from employing the insights derived from the dataset to manipulate consumers or engage in unethical business practices.

Additionally, as the dataset represents a snapshot of the AliExpress online retail platform at a specific time, it should not be used to make conclusive assertions about the overall consumer electronics market without corroboration from other sources.

Any other comments?

The AliExpress consumer electronics dataset provides a valuable resource for understanding the dynamics of this market segment on one of the world's largest e-commerce platforms. By carefully considering its strengths and limitations, users can effectively leverage the dataset to gain insights into consumer preferences, market trends, pricing strategies, and more.

However, responsible usage is crucial, and dataset consumers should combine this resource with domain expertise and other relevant information to make well-informed decisions. Regularly updating the dataset and expanding its coverage could further enhance its utility in driving innovation within the consumer electronics industry.