King Saud University
College of Computer and Information Sciences
Information Technology Department

جامعة الملك سعود
King Saud University
كلية علوم الحاسب والمعلومات
قسم تقنية المعلومات

DATA SCIENCE

Printer analysis

IT 362 phase 2

Prepared by:

Jana Aljomaih, 443200860
yasmen AL-suhaibani, 443201103
Ghaina Alhassnan, 443200495
Shaug Alyahya, 443200888
Jana AL-suhaibani, 443200428

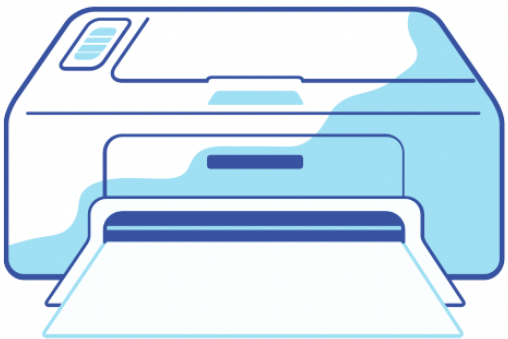

Table of Contents

A. Data Collection:

We've employed the Amazon Product Advertising API to collect printer data from the Amazon website. This API provides a key for requesting data, which we utilized to retrieve information about specific products or search results based on keywords. However, we encountered two primary challenges in this endeavor. Firstly, there's a restriction on the number of requests we can make, necessitating a subscription to access more data. Secondly, we faced difficulties retrieving information for certain products, particularly two of them, leading us to suspect that they may no longer be available.

- Bias and Fairness Section:

i. Summary of Your Research on Data Bias and Fairness

Research in data bias and fairness looks at how collecting, processing, and analyzing data the wrong way can lead to outcomes that aren't fair or correct. Bias can happen for many reasons, like how data is gathered, who or what is included in the data, where the data comes from, and how the data is analyzed. Being fair means making sure that the conclusions and decisions based on the data don't unfairly hurt or benefit any particular group or person.

ii. An Evaluation of Your Dataset's Potential Biases and Their Sources

When looking at our dataset about printers, we might find some biases, such as:

Selection Bias: If we only looked at certain brands or expensive printers, our dataset wouldn't show what's really out there in the whole printer market.

Availability Bias: We might have more information on some printers than others, making some more noticeable in our dataset.

Measurement Bias: If different sources report printer features differently, our data might not be consistent.

iii. The Implications of These Biases for the Fairness and Reliability of Your Project's Conclusions

Having biases in our dataset can make our conclusions about printers wrong. For example, if we only focus on certain brands, we might miss out on great options from other brands. This can mislead consumers, manufacturers, and stores. Also, if our data is biased, any predictions or analyses we do might not be reliable, making it harder to make good choices.

iv. Recommendations for Mitigating These Biases in Future Data collection and Analysis Efforts

To make our data better in the future, we could:

Expand Data Collection: Try to include more types of printers from different brands to get a fuller picture.

Standardize Data Measurement: Make sure everyone measures and reports about printers in the same way to keep things consistent.

Diversity in Data Sources: Get data from lots of different places to reduce the effect of bias from any one source.

Bias Analysis: Regularly check for and fix biases in our data.
Transparent Methodology: Be open about how we collect and analyze data so others can help make it better.

By doing these things, we can work towards making our data analysis fairer, more accurate, and more reliable.

B. Data Processing and Cleaning Section:

The process of dataset cleaning started with discovering and understanding the dataset dimensions, after that we applied the following procedure:

- ## Removing useless characters:

As attribute **'Title'** is in natural language form, it included unnecessary additional characters to the dataset that could come against our processing approach, for that we removed special characters as a first step followed by removing and generalizing the word "Email" in the attribute **'Special feature',** then we cleaned out the **'Brand'** names from the additional charaters coming from data collecting 'â\x80\x8e' by that we made sure that Brand names are valid and accurate. After understanding the inputs of special feature attribute which could be either: monochrome or color, our third step in this task was removing any additional number from **'Printer output'** for a neat and clear dataset.

- ## Dealing with nulls:

In this task we first viewed the nulls using a heatmap we found out that attributes: **Controller** type (587/593nulls), **Protection plans** (336/593 nulls), **variants** (246/593 nulls), and **Maximum print speed monochrome** (159/593 nulls) hold a great number of nulls. Surely, we didn't rush with the repairing plan! we held arranged meetings with the team members and had a very clear and secure discussions to the conclusion that three out of previous four columns are insignificant to the dataset, question's answers, and both classification and clustering. Those three columns are: **Controller type, Protection plans**, and **variants**. Yet we kept **Maximum print speed color** and **Maximum print speed monochrome** as they complete each other: if the printer speed is color then the number of speed is provided as input under that column, if the speed is monochrome then number of speed is provided as input under the later column. For that we will fill in the empty nulls in another task described later.

- ## Changing type of series:

Rating and **Rating total** were converted from object to float, **Maximum print speed color** and **Maximum print speed** monochrome underwent the same process after removing the unit (ppm)

We have realized that **weight** shall be changed from object to float too, but as it contains different units: gram, kilogram, ounce, and pound. We will have to push the changing of attribute type till we apply data transformation.

- ## Dealing with duplicate rows and empty entities:

To address the right and true number of empty input we used the following code snippet:

round((df.isin([' ',np.nan,'NULL']).sum()) , 0)

We will walk through these columns one by one:

Bestsellers_rank was filled with zeros, explanation states that rank 1 is considered the best rank a printer could have and it decreases in ascending order for that we assigned value 0 to non ranke d printers assuring 0 is out of ranking sector not to cause faulty in other printers rank and yet it i s considered a good option as it wont push us to hold back from dealing with this column as float .

Brand underwent long consideration and studying the needs of the dataset, for the plan to come out as the 4 null values in brand were dropped as they can't be detected.

Price and **Rating** were filled out using the groupby according to the brand they belong to, good explanation is that Brand hold a known range of prices and reputation -if a brand is known for good printers it is highly anticipated that its new printer will be in the same rating perimeter as it undergoes same well defined test established by the Brand company- we used this information to calculate the mean of each brand price and rating and if a printer comes from a unique brand to the dataset then it is filled with the overall mean.

Ratings total cannot be filled using the mean, as we did with price or rating, because it depends on the number of users who have rated the printer. We will input a value of 1 into null cells, ensuring that at least one user has provided a rating for the printer, as indicated by the rating column.

Maximum print speed color and **Max print speed monochrome** had dependency on each other.

So if the speed color is null, fill the speed with zero (no speed), as it surely uses monochrome. and if monochrome is empty, then fill the speed with zero, as it surely is color. We made sure that we didn't fill a printer with zeros in both columns by using the following for loop:

for entity in Maximum print speed monochrome :

if Maximum print speed monochrome doesn't equal zero:

then Printer output equals monochrome

if Maximum print speed color doesn't equal zero:

then Printer output equals monochrome

So if the **Printer output** nulls are gone, then no duplicate zeros happened.

Checking the nulls in the dataset, we see that all nulls of the previously mentioned attributes have been filled successfully:

ï»¿Asin	0
Brand	0
price	0
Bestsellers_rank	0
Rating	0
Ratings_total	0
Printer_Output	0
Maximum_Print_Speed_Color	0
Max_Printspeed_Monochrome	0

- ## Data Transformation:

For **weight,** we have generalized gram as the desired unit, then after converting all input to grams, we delered the unit to make sure the column is float.

Dimensions have been split into three different columns: height, length, and depth, then we filled the nulls with the mean and converted the attributes to numeric.

Newer model has been converted to the binary inputs 0 (not newer model) and 1 (newer model).

Connectivity technology was dealt with after looking deeply into the dataset, as we found out that connectivity technology could be: bluetooth, wi-fi, usb, ethernet, wired, wireless, manual, or a combination of the above.

We first lowercased the column inputs, then in a for loop, if the value has either of the above words, keep it as it is, or else replace it with the mode bluetooth.

We used this method instead of filnull, as some of the not-null entities contain false input.

Then we cleaned the rest of the inputs and removed additional spaces from both the left and right sides.

Special feature: In the subsequent steps, any missing values within the column will be populated with the feature descriptor 'print,' denoting a feature commonly associated with printers. Following the imputation process, a text processing function will be applied to transform the textual data into an array format within the respective column. Consequently, a

new column will be created to enumerate the elements , thereby reflecting the total count of features attributed to a printer.

Printing technology: in summary, a replacement of empty strings to np.nan was done to ensure uniformity, then for each brand, we determined the most common printing technology and filled any remaining missing value with ‘other’. Here are the detailed steps:

Standardization of Missing Values: Initially, we addressed the representation of missing data by converting all empty strings within the "Printing_Technology" column to np.nan, establishing a uniform marker for missing values. This step is crucial for the consistency and reliability of subsequent imputation methods.

Group-wise Mode Imputation: Leveraging the assumption that products from the same brand are likely to share common printing technologies, we performed mode imputation within groups defined by the "Brand" column. This strategy allowed us to fill missing "Printing_Technology" values with the most frequent (mode) value within each brand, reflecting a nuanced understanding of the dataset's domain.

Fallback Strategy for Unimputable Groups: For brands where a mode could not be determined (due to all values being missing or no clear mode), we applied a fallback strategy by filling the remaining missing values with a default placeholder ('other'). This step ensured no missing values remained in the "Printing_Technology" column, maintaining the dataset's integrity for analysis or modeling without the complications of missing data.

Verification of Data Completeness: The final step involved verifying that no missing values remained in the "Printing_Technology" column after the imputation process. This step is essential to confirm the effectiveness of the imputation and the readiness of the dataset for further use.

steps

Color: In summary, we cleaned the inputs, then applied the same method of **Printing technology**: replacement of empty strings to np.nan was done to ensure uniformity, then for each brand, we determined the most common printing technology and filled any remaining missing value with ‘other’.

- ## Generalizing column names:

For a neat and presentable dataset, we converted the column names into:

Asin
Brand
Title
Price
keywords
Categories
Bestsellers_rank
Color
Rating
Ratings_total
Bundles
Frequently_bought_together
Newer_model
Connectivity_Technology
Printing_Technology
Special_Feature
Printer_Output
Maximum_Print_Speed_Color
Max_Printspeed_Monochrome
Weight
Depth
Width
Height

- ## Key Concepts Employed:

Data Cleaning and Imputation: Fundamental tasks in data preprocessing that enhance the quality and usability of the dataset.

Group-wise Imputation: A strategy that uses categorical groupings within the data to make informed imputation decisions based on common characteristics or behaviors within those groups.

Domain Knowledge Application: The imputation method was chosen with an understanding of the dataset's specific context, utilizing knowledge about the likely similarity of printing technologies within brands.

Ensuring Completeness and Integrity: A critical aspect of data preprocessing is ensuring that the dataset is free of missing values in key columns to avoid issues in subsequent analyses or modeling stages.

This process exemplifies a thoughtful approach to handling missing data, tailored to the dataset's specific characteristics and the analytical goals at hand.

C. Exploratory Data Analysis (EDA):

In conducting an exploratory data analysis (EDA) of a dataset comprising information on 424 printers, a comprehensive approach was taken to uncover patterns, trends, and anomalies within the data. This analysis utilized Python as the primary programming language, leveraging powerful libraries such as Pandas for data manipulation, Matplotlib and Seaborn for visualization, and NumPy for numerical computations. The EDA encompassed various types of analysis, including statistical summaries, correlation analysis, and visual explorations, to provide a multi-faceted view of the dataset.

- ## Statistical Summaries:

Key numerical insights revealed a wide range of printer prices with an average of approximately 99.37, suggesting a market with diverse offerings catering to different budget segments. The printers were associated with a moderate number of keywords on average (around 18), indicating varied product features and marketing focuses. Interestingly, the dataset showed a significant variance in sales performance (bestsellers rank) and a high average rating of 4.32 out of 5, highlighting general consumer satisfaction. Anomalies included a large discrepancy between the number of newer and older model printers, with only 24 considered newer models, hinting at a potential market for technological upgrades or a consumer preference for established models.

- ## Visualizations:

Through charts and graphs, the analysis provided a visual representation of the data, revealing:

- A concentration of lower-priced printers, indicating a market trend towards affordability.

- A general clustering of printer ratings is around 4 stars, underscoring widespread consumer satisfaction.

- Brand dominance by HP, followed by significant presences by Canon and Brother, which might influence market strategy and inventory decisions.

- A lack of a clear, strong correlation between printer prices and ratings, suggesting that higher prices do not necessarily equate to higher consumer satisfaction.

- ### Correlation Analysis:

A heatmap analysis indicated minimal strong linear relationships among most dataset attributes, with a few exceptions like the positive correlation between 'Weight' and 'Depth', and between 'Width' and 'Height'. This suggests that physical dimensions and possibly shipping considerations might influence product design and packaging strategies.

- ### Initial Insights:

- The market is characterized by a broad price range, with a focus on more affordable, accessible options for the broader consumer base.

- Despite the variety in price and features, consumer satisfaction remains consistently high, highlighting quality and performance as key factors over price.

- The strong presence of established brands and the slow introduction of newer models suggest brand loyalty and a potential hesitancy towards adopting newer technologies among consumers.

- The lack of strong correlations between many attributes indicates a complex interplay of factors influencing consumer choice, beyond just price or specific features.

This EDA provided valuable insights into the printer market, revealing the complexity of consumer preferences and market dynamics. It underscores the importance of considering a wide range of factors, from brand reputation and product features to price and consumer reviews, in understanding market trends and making informed decisions in product development, marketing, and sales strategies.

D. Future Steps Section:

For future steps based on the EDA findings, the project could involve conducting statistical tests to understand the relationship between variables such as price and printer ratings or print speed and printing technology. Creating advanced visualizations would help in identifying trends and correlations in the data. To deepen the analysis, additional data related to printer sales, customer reviews, or after-sales services could be gathered. Predictive modeling could be employed to forecast printer ratings based on features like price, technology, and dimensions, and to predict market trends. Comparative studies could also be conducted to identify market leaders and understand trends in printer technology and consumer preferences. Delving deeper into the EDA results will provide specific insights to guide these steps. The potential next steps for our project include:

• Hypotheses for further investigation:

- Explore the impact of brand loyalty on ratings, especially with brands like HP or Canon.
- Assess the impact of newer printing technologies on ratings.
- Examine how different market segments influence ratings and sales.
- Determine if promotional discounts positively impact ratings and sales.
- Assess whether printers with more features tend to receive higher ratings.

These hypotheses aim to deepen understanding of the factors that influence customer perceptions, sales performance, and market dynamics within the printer industry.

• Additional data that could be collected:

Gathering a comprehensive set of data is crucial for gaining deeper insights into various aspects of the printer industry. This includes customer demographics, preferences, and behavior, as well as competitor information such as products, pricing strategies, and market share. Detailed product features and specifications, along with promotional activities and marketing strategies, also play a significant role. Customer satisfaction surveys, industry reports, and market trends provide valuable context, while environmental impact and sustainability metrics offer insight into emerging consumer concerns. Additionally, warranty and support information are essential for understanding customer service dynamics. Online and social media data can be analyzed for sentiment analysis and user-generated content, providing real-time feedback and consumer sentiment.

• Model Development:

Build predictive models to understand factors influencing customer ratings or sales performance. Possible models include:

- Regression models predict ratings based on price, printing speed, weight, features, etc.
- Classification models to predict whether a printer will be a bestseller or not, whether the rating will be high or not, type of connectivity technology, type of printing technology, and price.
- Clustering algorithms to identify similarities and differences across distinct market segments based on price, features, and ratings.