Problem Statement

Analysis of laptops across various platforms to understand the factors influencing the price and specifications influencing customers interest in purchasing a laptop.

Through this analysis we can find out the elements that are important in fixing the prices, and the elements that makes the major contribution in the price. Also finding the main features of the laptop that are prior requirements of the customers , and which are making a clear impression on the product to the customer.

This analysis will help both seller and customer to obtain the best beniefits as for seller the information about the requirements and the first things that customer observe while before the buying the laptop will help making better prices and better products according to the customers requirements.For customers it would be helpful in chossing the right product acording to their price range and the best features that can be obtained.

For this we decieded to obtain data from online ecommerce weebsites Amazon,Flipkart,BestBuy

Data Collection and Cleaning - Flipkart

- The dataset was scrapped from flipkart website.
- Scraping and Cleaning done by:
  - Name: Vaibahv Saran
  - UB ID: 50615031

Data Collection - Flipkart

1. Importing Modules

```
import requests
from bs4 import BeautifulSoup
import re
import numpy as np
import pandas as pd
import time
import random
```

Explanation for Modules

- `requests` module is to send a request to the URL to fetch the data.

- `BeautifulSoup` is a class using the object of which we will deal with the scraped HTML data.
- `re` is for using regex patterns to filter out data and create our dataframe in an organized format.
- `numpy` and `pandas` module if for manipulating data values and handling the data overall.
- `time` module is used to create a time delay during scraping.
- `random` module is used to generate random numbers to be used during scraping time delays.

2. Scraping All The Webpages Of Flipkart For Laptop Data

image.jpg

- As highlighted in the red box of the above image we have access to **68 pages** of flipkart to scrape the data from.
- So based on that we will write the code to scrape the data

```
# Defining Request Headers to scrape the data
request_header = {
    'Content-Type': 'text/html; charset=UTF-8',
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:109.0)
Gecko/20100101 Firefox/119.0',
    'Accept-Encoding': 'gzip, deflate, br',
    'Referer': 'https://www.flipkart.com/',
    'Origin': 'https://www.flipkart.com',
    'Accept-Language': 'en-US,en;q=0.9'
}
```

Understanding The Request Headers

- **Content-Type**: This tells the server what kind of data you're sending. In this case, it's HTML text with a specific character set (UTF-8).
- **User-Agent**: This identifies the browser and operating system making the request. It helps the server understand how to format the response. For example, we are using Firefox on a Windows 10 machine.
- **Accept-Encoding**: This tells the server which compression methods your client can handle. Here, it indicates that the client can accept responses compressed with gzip, deflate, or br (Brotli).
- **Referer**: This indicates the URL from which the request originated. It helps the server understand the context of the request. In this scenario, it shows that the request is coming from Flipkart's website.
- **Origin**: Similar to Referer, this specifies the origin of the request, which is also Flipkart in this case. It's used for security purposes, particularly with cross-origin requests.
- **Accept-Language**: This tells the server which languages your client prefers. Here, it indicates a preference for US English, but can also accept other forms of English.

```
# Storing the URL as a f-string
page = 1
URL = f"https://www.flipkart.com/search?
```

```
q=laptop&otracker=search&otracker1=search&marketplace=FLIPKART&as-show=on&as=off&page={page}"
```

- The URL is stored as f-string becuase, by changing the page number in the URL, we can access the next page data, this can be utilized in conjunction with for loop to scrape all the available data.

```
# Scraping Code

total_pages = 68 # Total number of pages being scraped
i = 1 # Counter to self verify the pages being scraped successfully
raw_text = [] # List to store all the raw html code

# Loop to iterate over all the pages by changing the f-string URL
for page in range (1, total_pages+1):

    # Fetching the data from URL based on the above request headers
    response = requests.get(URL, headers=request_header)

    # Random number to be used as time delay in order to make the script behaviour more human like
    delay = random.randint(5,10)
    print("Time Delay:",delay,end=" seconds    : ")

    # While Loop: covers the edge case wherein the first attempt to fetch the data failed,
    # by continuously requesting the data at irregular time intervals in order to mimic human behavior
    while response.status_code!=200:
        time.sleep(delay)
        response = requests.get(URL,headers=request_header)

    # Confirmation Message of Successful Scrape
    print("Page",i," status:",response)

    # Incrementing Page Counter
    i+=1

    # Appending the raw HTML code in the list
    raw_text.append(response.text)

    # A random delay before requesting the data from next page
    time.sleep(delay)
```

```
Time Delay: 9 seconds    : Page 1  status: <Response [200]>
Time Delay: 7 seconds    : Page 2  status: <Response [200]>
Time Delay: 9 seconds    : Page 3  status: <Response [200]>
Time Delay: 9 seconds    : Page 4  status: <Response [200]>
Time Delay: 6 seconds    : Page 5  status: <Response [200]>
Time Delay: 6 seconds    : Page 6  status: <Response [200]>
```

```
Time Delay: 7 seconds    : Page 7  status: <Response [200]>
Time Delay: 9 seconds    : Page 8  status: <Response [200]>
Time Delay: 8 seconds    : Page 9  status: <Response [200]>
Time Delay: 10 seconds    : Page 10  status: <Response [200]>
Time Delay: 9 seconds    : Page 11  status: <Response [200]>
Time Delay: 7 seconds    : Page 12  status: <Response [200]>
Time Delay: 9 seconds    : Page 13  status: <Response [200]>
Time Delay: 7 seconds    : Page 14  status: <Response [200]>
Time Delay: 5 seconds    : Page 15  status: <Response [200]>
Time Delay: 7 seconds    : Page 16  status: <Response [200]>
Time Delay: 10 seconds    : Page 17  status: <Response [200]>
Time Delay: 8 seconds    : Page 18  status: <Response [200]>
Time Delay: 5 seconds    : Page 19  status: <Response [200]>
Time Delay: 8 seconds    : Page 20  status: <Response [200]>
Time Delay: 6 seconds    : Page 21  status: <Response [200]>
Time Delay: 5 seconds    : Page 22  status: <Response [200]>
Time Delay: 9 seconds    : Page 23  status: <Response [200]>
Time Delay: 10 seconds    : Page 24  status: <Response [200]>
Time Delay: 7 seconds    : Page 25  status: <Response [200]>
Time Delay: 8 seconds    : Page 26  status: <Response [200]>
Time Delay: 10 seconds    : Page 27  status: <Response [200]>
Time Delay: 8 seconds    : Page 28  status: <Response [200]>
Time Delay: 5 seconds    : Page 29  status: <Response [200]>
Time Delay: 10 seconds    : Page 30  status: <Response [200]>
Time Delay: 7 seconds    : Page 31  status: <Response [200]>
Time Delay: 6 seconds    : Page 32  status: <Response [200]>
Time Delay: 9 seconds    : Page 33  status: <Response [200]>
Time Delay: 8 seconds    : Page 34  status: <Response [200]>
Time Delay: 9 seconds    : Page 35  status: <Response [200]>
Time Delay: 8 seconds    : Page 36  status: <Response [200]>
Time Delay: 5 seconds    : Page 37  status: <Response [200]>
Time Delay: 5 seconds    : Page 38  status: <Response [200]>
Time Delay: 8 seconds    : Page 39  status: <Response [200]>
Time Delay: 8 seconds    : Page 40  status: <Response [200]>
Time Delay: 5 seconds    : Page 41  status: <Response [200]>
Time Delay: 6 seconds    : Page 42  status: <Response [200]>
Time Delay: 9 seconds    : Page 43  status: <Response [200]>
Time Delay: 7 seconds    : Page 44  status: <Response [200]>
Time Delay: 6 seconds    : Page 45  status: <Response [200]>
Time Delay: 7 seconds    : Page 46  status: <Response [200]>
Time Delay: 10 seconds    : Page 47  status: <Response [200]>
Time Delay: 6 seconds    : Page 48  status: <Response [200]>
Time Delay: 7 seconds    : Page 49  status: <Response [200]>
Time Delay: 7 seconds    : Page 50  status: <Response [200]>
Time Delay: 10 seconds    : Page 51  status: <Response [200]>
Time Delay: 5 seconds    : Page 52  status: <Response [200]>
Time Delay: 10 seconds    : Page 53  status: <Response [200]>
Time Delay: 10 seconds    : Page 54  status: <Response [200]>
Time Delay: 5 seconds    : Page 55  status: <Response [200]>
```

```
Time Delay: 5 seconds    : Page 56  status: <Response [200]>
Time Delay: 9 seconds    : Page 57  status: <Response [200]>
Time Delay: 9 seconds    : Page 58  status: <Response [200]>
Time Delay: 10 seconds    : Page 59  status: <Response [200]>
Time Delay: 8 seconds    : Page 60  status: <Response [200]>
Time Delay: 7 seconds    : Page 61  status: <Response [200]>
Time Delay: 5 seconds    : Page 62  status: <Response [200]>
Time Delay: 9 seconds    : Page 63  status: <Response [200]>
Time Delay: 8 seconds    : Page 64  status: <Response [200]>
Time Delay: 9 seconds    : Page 65  status: <Response [200]>
Time Delay: 6 seconds    : Page 66  status: <Response [200]>
Time Delay: 10 seconds    : Page 67  status: <Response [200]>
Time Delay: 10 seconds    : Page 68  status: <Response [200]>
```

3. Saving The Raw HTML Data in CSV

- Now we will save the raw HTML code for each page in a CSV by converting the list into a dataframe.
- Saving in CSV will ensure that we don't have to scrape the entire data everytime we want to work on the data as scraping itself is a time consuming process.

```
# Converting the list to Data Frame
df = pd.DataFrame(raw_text,columns=["Raw Data"])

# Printing a sample to ensure correct data format
df.head()
```

```
                                                Raw Data
0  <!doctype html><html lang="en"><head><link hre...
1  <!doctype html><html lang="en"><head><link hre...
2  <!doctype html><html lang="en"><head><link hre...
3  <!doctype html><html lang="en"><head><link hre...
4  <!doctype html><html lang="en"><head><link hre...
```

```
# Dataframe has been created successfully and can now be saved in a
CSV file
df.to_csv(r"data\raw.csv")
```

Data Cleaning - Flipkart

1. Importing Requisite Libraries

```
import numpy as np
import pandas as pd
import re
from bs4 import BeautifulSoup
```

List Of The Features To Be Extracted From The Raw Data

- laptop_company = []

- processor_company = []
- processor = []
- operating_system = []
- RAM = []
- storage = []
- storage_type = [] # If not SSD Default will be HDD
- rating = []
- No_reviews = []
- screen_size = []
- price = [] # Target column

1.1 Filtering Features From Raw Data

```
# Loading the raw CSV
df = pd.read_csv(r"data\raw.csv")
df.head()
```

```
   Unnamed: 0                                               Raw Data
0           0  <!doctype html><html lang="en"><head><link hre...
1           1  <!doctype html><html lang="en"><head><link hre...
2           2  <!doctype html><html lang="en"><head><link hre...
3           3  <!doctype html><html lang="en"><head><link hre...
4           4  <!doctype html><html lang="en"><head><link hre...
```

```
pages = []
for i,j in df.iterrows():
    pages.append(j["Raw Data"])

# Iterating over all pages and for each page filtering out the laptop
brand for each product
laptop_brand = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="tUxRFH"):
        regex = re.findall("Compare(\w+)",i.text)
        if regex:
            laptop_brand.append(regex[0])
        else:
            laptop_brand.append(np.nan)

len(laptop_brand)
```

```
1632
```

```
# Filtering out laptop names
laptop_name = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):
        regex = re.findall("(^.+)\s(?:Intel|intel|AMD|M1|M2|M3|
```

```
Chromebook|Snapdragon)",i.text)
        if regex:
            laptop_name.append(regex[0])
        else:
            laptop_name.append(np.nan)

len(laptop_name)
```

```
1632
```

```
processor = []
processor_company = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):

        # Regex to find the processor company of the laptop
        regex1 = re.findall("Intel|intel|AMD|M1|M2|M3|Chromebook|
Snapdragon",str(i.text))
        if regex1:
            processor_company.append(regex1[0])
        else:
            processor_company.append(np.nan)

        # Regex to find the exact processor in the laptop
        regex2 = re.findall("(?:Intel|intel|AMD|M1|M2|M3|Chromebook|
Snapdragon)\s(.+) - ",str(i.text))
        if regex2:
            processor.append(regex2[0])
        else:
            processor.append(np.nan)

# Checking the number of datapoints after applying the regex
print(len(processor_company))
```

```
1632
```

```
print(len(processor))
```

```
1632
```

```
# Filtering out operating System of the laptops
operating_system = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):

        # Regex to find Operating Systems
        regex = re.findall(".+((?:Windows 10|Mac OS|DOS|Andorid|
Chrome|Windows 11|Windows 11 Home))",str(i.text))
```

```
        if regex:
            operating_system.append(regex[0])
        else:
            operating_system.append(np.nan)

# Checking the number of datapoints
len(operating_system)
```

```
1632
```

```
# Filtering out RAM of the laptops
RAM = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):

        # Regex to find RAM
        regex = re.findall("(\d+)\sGB\/",str(i.text))

        if regex:
            RAM.append(regex[0])
        else:
            RAM.append(np.nan)

len(RAM)
```

```
1632
```

```
# Filtering out Storage of the laptops
storage = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):

        # Regex to find Storage Size
        regex = re.findall("\d+\sGB\/(\d+)\s(?:GB|TB)\s(?:SSD|HDD|
EMMC)",str(i.text))

        if regex:
            storage.append(regex[0])
        else:
            storage.append(np.nan)

len(storage)
```

```
1632
```

```
# Filtering out Storage type of the laptops
storage_type = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="KzDlHZ"):
```

```
        # Regex to find Storage type
        regex = re.findall("\d+\sGB\/\d+\s(?:GB|TB)\s((?:SSD|HDD|
EMMC))",str(i.text))

        if regex:
            storage_type.append(regex[0])
        else:
            storage_type.append(np.nan)

len(storage_type)
```

```
1632
```

```
# Filtering out Rating of the laptops
rating = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="tUxRFH"):

        r = i.find("div",class_="XQDdHH")

        if r:
            rating.append(r.text)
        else:
            rating.append(np.nan)

len(rating)
```

```
1632
```

```
# Filtering out number of reviews for each laptops
No_reviews = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="tUxRFH"):
        p =i.find("span",class_="Wphh3N")
        if p:
            regex = re.findall("\&\s(.+)\sReviews",p.text)
            if regex:
                No_reviews.append(regex[0])
            else:
                No_reviews.append(np.nan)
        else:
            No_reviews.append(np.nan)

len(No_reviews)
```

```
1632
```

```
# Filtering out Screen Size of the laptops
screen_size = []
```

```
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="_6NESgJ"):
        regex = re.findall("\d+?\.?\d+?",i.text)
        if regex:
            screen_size.append(regex[0])
        else:
            screen_size.append(np.nan)

len(screen_size)
```

```
1632
```

```
# Filtering out the target column: price of the laptop
price = []
for page in pages:
    soup = BeautifulSoup(page)
    for i in soup.find_all("div",class_="tUxRFH"):

        p =i.find("div",class_="Nx9bqj _4b5DiR")

        if p:
            price.append(p.text)
        else:
            price.append(np.nan)

len(price)
```

```
1632
```

```
# Creating a dataframe from all the extracted features present in list
feature_dict = {"Laptop_Brand":laptop_brand,
                "Laptop_Name":laptop_name,
                "Processor_Company":processor_company,
                "Processor":processor,
                "Operating_System":operating_system,
                "RAM":RAM,
                "Storage":storage,
                "Storage_Type":storage_type,
                "Screen_Size":screen_size,
                "Rating":rating,
                "Number_of_Reviews": No_reviews,
                "Price":price}

# Creating a dataframe from the above dictionary
laptop_df = pd.DataFrame(feature_dict)

# Saving the dataframe as a csv file for further analysis
laptop_df.to_csv(r"data\flipkart_laptop_data.csv",index=False)
```

Data Cleaning - Flipkart

1. Loading the Necessary Modules

```
import numpy as np
import pandas as pd
import re
import warnings
warnings.filterwarnings("ignore")
```

2. Loading the CSV file

```
laptop_df = pd.read_csv(r"data\flipkart_laptop_data.csv")
laptop_df.head()
```

```
  Laptop_Brand  Laptop_Name Processor_Company                Processor
\
0           HP    HP Victus             Intel         Core i5 12th Gen

1          MSI  MSI Thin 15             Intel  Core i5 12th Gen 12450H

2           HP    HP Laptop               AMD  Ryzen 3 Quad Core 5300U

3         Acer     Acer One             Intel  Core i3 11th Gen 1115G4

4           HP           HP               AMD  Ryzen 5 Hexa Core 5500U


  Operating_System  RAM  Storage Storage_Type  Screen_Size  Rating  \
0       Windows 11   16      512          SSD           12     4.4
1       Windows 11   16      512          SSD           12     4.3
2       Windows 11    8      512          SSD           11     4.3
3       Windows 11    8      512          SSD           11     4.2
4       Windows 11   16      512          SSD           16     4.3

   Number_of_Reviews    Price
0               38.0  ₹58,990
1               34.0  ₹57,990
2              482.0  ₹30,999
3              571.0  ₹26,990
4              268.0  ₹42,990
```

3. Cleaning of Dataset

```
laptop_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1632 entries, 0 to 1631
Data columns (total 12 columns):
```

```
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Laptop_Brand       1632 non-null   object
 1   Laptop_Name        1632 non-null   object
 2   Processor_Company  1632 non-null   object
 3   Processor          1632 non-null   object
 4   Operating_System   1632 non-null   object
 5   RAM                1632 non-null   int64
 6   Storage            1632 non-null   int64
 7   Storage_Type       1632 non-null   object
 8   Screen_Size        1632 non-null   int64
 9   Rating             1503 non-null   float64
 10  Number_of_Reviews  1503 non-null   float64
 11  Price              1623 non-null   object
dtypes: float64(2), int64(3), object(7)
memory usage: 153.1+ KB
```

Observation

- There are a total of `1632` datapoints but looking at the columns `Rating` and `Number_of_Reviews`, there are some null values which needs to be dealt with later.
- Need to check for wrong values or outliers in the data.
- The target column `Price` should be integer but is stored as an object so it must be converted to right datatype as well as the missing data needs to be replaced.

```
# Inspecting Price Column
laptop_df["Price"].head(10)
```

```
0    ₹58,990
1    ₹57,990
2    ₹30,999
3    ₹26,990
4    ₹42,990
5    ₹64,990
6    ₹52,990
7    ₹35,990
8    ₹20,990
9    ₹36,990
Name: Price, dtype: object
```

Observation

- Based on the above cell output, we see that Price is being treated as object column becuase of an extra symbol `₹` and `,`.
- Therefore they need to be removed as they dont contribute in the analysis.

```
# Removing extra characters from the price column
laptop_df["Price"] =
laptop_df["Price"].str.replace(',','').str.replace('₹', '')
laptop_df.head()
```

```
  Laptop_Brand  Laptop_Name Processor_Company                Processor
\
0           HP    HP Victus             Intel         Core i5 12th Gen

1          MSI  MSI Thin 15             Intel  Core i5 12th Gen 12450H

2           HP    HP Laptop               AMD  Ryzen 3 Quad Core 5300U

3         Acer     Acer One             Intel  Core i3 11th Gen 1115G4

4           HP           HP               AMD  Ryzen 5 Hexa Core 5500U


  Operating_System  RAM  Storage Storage_Type  Screen_Size  Rating  \
0       Windows 11   16      512          SSD           12     4.4
1       Windows 11   16      512          SSD           12     4.3
2       Windows 11    8      512          SSD           11     4.3
3       Windows 11    8      512          SSD           11     4.2
4       Windows 11   16      512          SSD           16     4.3

   Number_of_Reviews  Price
0               38.0  58990
1               34.0  57990
2              482.0  30999
3              571.0  26990
4              268.0  42990
```

```
# Analysing the complete description summary of the dataframe
laptop_df.describe(include="all").T
```

```
                    count unique                     top  freq
mean  \
Laptop_Brand         1632      9                      HP   388
NaN
Laptop_Name          1632     49                      HP   136
NaN
Processor_Company    1632      4                   Intel  1040
NaN
Processor            1632     38  Core i3 12th Gen 1215U   182
NaN
Operating_System     1632      2              Windows 11  1541
NaN
RAM                1632.0    NaN                     NaN   NaN
12.252451
Storage            1632.0    NaN                     NaN   NaN
440.907475
Storage_Type         1632      2                     SSD  1541
NaN
Screen_Size        1632.0    NaN                     NaN   NaN
19.865809
```

```
Rating              1503.0    NaN                    NaN   NaN
4.203127
Number_of_Reviews   1503.0    NaN                    NaN   NaN
246.401863
Price                 1623     53                  54990   100
NaN

                           std   min    25%    50%    75%    max
Laptop_Brand               NaN   NaN    NaN    NaN    NaN    NaN
Laptop_Name                NaN   NaN    NaN    NaN    NaN    NaN
Processor_Company          NaN   NaN    NaN    NaN    NaN    NaN
Processor                  NaN   NaN    NaN    NaN    NaN    NaN
Operating_System           NaN   NaN    NaN    NaN    NaN    NaN
RAM                   5.865025   4.0    8.0    8.0   16.0   32.0
Storage             171.443332   1.0  512.0  512.0  512.0  512.0
Storage_Type               NaN   NaN    NaN    NaN    NaN    NaN
Screen_Size          18.555163  11.0   11.0   12.0   16.0   81.0
Rating                0.238804   3.3    4.1    4.2    4.3    5.0
Number_of_Reviews   202.948019   0.0   25.0  214.0  467.0  597.0
Price                      NaN   NaN    NaN    NaN    NaN    NaN
```

Observation Based Tasks:

- Convert the price column to int after handling missing values.
- Convert the laptop names to proper product names
- In the storage column, there is a min vlaue of 1, which is measuring the data in TB, so all the values must be converted to a common GB measurement.
- Check the screen size for values and impute the outliers accordingly.
- In rating and no of reviews replace null values with 0.

```
# Analyzing the datapoints which have null price
laptop_df[laptop_df["Price"].isnull()]

     Laptop_Brand Laptop_Name Processor_Company
Processor  \
25         Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
145        Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
217        Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
241        Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
265        Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
641        Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
1457       Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX
1553       Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
```

```
13450HX
1577       Lenovo  Lenovo LOQ             Intel  Core i5 13th Gen
13450HX

     Operating_System  RAM  Storage Storage_Type  Screen_Size  Rating
\
25         Windows 11   16      512          SSD           13     4.2

145        Windows 11   16      512          SSD           13     4.2

217        Windows 11   16      512          SSD           13     4.2

241        Windows 11   16      512          SSD           13     4.2

265        Windows 11   16      512          SSD           13     4.2

641        Windows 11   16      512          SSD           13     4.2

1457       Windows 11   16      512          SSD           13     4.2

1553       Windows 11   16      512          SSD           13     4.2

1577       Windows 11   16      512          SSD           13     4.2


      Number_of_Reviews Price
25                 44.0   NaN
145                44.0   NaN
217                44.0   NaN
241                44.0   NaN
265                44.0   NaN
641                44.0   NaN
1457               44.0   NaN
1553               44.0   NaN
1577               44.0   NaN
```

Observation

- All the above data points are duplicated, so will be dropped.
- The process of handling will be as follows:
  - Removing all the null values in Price column excpet for the 1 datapoint.
  - Replacing the null value in Price with the mean of 50th percentile and 75th percentile of all Lenovo Laptop Prices.

```
# Droping 8 rows where Laptop_Name is 'Lenovo LOQ' and Price is null
lenovo_loq_null_price = laptop_df[(laptop_df['Laptop_Name'] == 'Lenovo
LOQ') & (laptop_df['Price'].isnull())]
laptop_df = laptop_df.drop(lenovo_loq_null_price.index[:8])

# Calculating the mean of the 50th and 75th percentiles of the Price
column for Lenovo laptops
```

```
lenovo_prices = laptop_df[laptop_df['Laptop_Brand'] == 'Lenovo']
['Price'].dropna().astype(float)
percentile_50 = lenovo_prices.quantile(0.50)
percentile_75 = lenovo_prices.quantile(0.75)
mean_price = (percentile_50 + percentile_75) / 2

# Replacing the null value in the remaining 'Lenovo LOQ' row with the
calculated mean price
laptop_df.loc[(laptop_df['Laptop_Name'] == 'Lenovo LOQ') &
(laptop_df['Price'].isnull()), 'Price'] = mean_price

# Verifying the changes
laptop_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1624 entries, 0 to 1631
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Laptop_Brand       1624 non-null   object
 1   Laptop_Name        1624 non-null   object
 2   Processor_Company  1624 non-null   object
 3   Processor          1624 non-null   object
 4   Operating_System   1624 non-null   object
 5   RAM                1624 non-null   int64
 6   Storage            1624 non-null   int64
 7   Storage_Type       1624 non-null   object
 8   Screen_Size        1624 non-null   int64
 9   Rating             1495 non-null   float64
 10  Number_of_Reviews  1495 non-null   float64
 11  Price              1624 non-null   object
dtypes: float64(2), int64(3), object(7)
memory usage: 164.9+ KB
```

```
# Converting Price column to numbers
laptop_df["Price"] = laptop_df["Price"].astype("float")
laptop_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 1624 entries, 0 to 1631
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Laptop_Brand       1624 non-null   object
 1   Laptop_Name        1624 non-null   object
 2   Processor_Company  1624 non-null   object
 3   Processor          1624 non-null   object
 4   Operating_System   1624 non-null   object
 5   RAM                1624 non-null   int64
 6   Storage            1624 non-null   int64
```

```
 7   Storage_Type       1624 non-null   object
 8   Screen_Size        1624 non-null   int64
 9   Rating             1495 non-null   float64
 10  Number_of_Reviews  1495 non-null   float64
 11  Price              1624 non-null   float64
dtypes: float64(3), int64(3), object(6)
memory usage: 164.9+ KB
```

```
# Removing the brand name from Laptop Name to just have product name
def remove_company(name):
    words = name.split(' ',1)

    if len(words)>1:
        # Removing anything which comes after , or -
        trunc = words[1].split(',')[0].split('-')[0].strip()
        return trunc if trunc else name

    return name

laptop_df["Laptop_Name"] =
laptop_df["Laptop_Name"].apply(remove_company)

laptop_df.head(10)
```

```
  Laptop_Brand  Laptop_Name Processor_Company                Processor
\
0           HP       Victus              Intel         Core i5 12th Gen

1          MSI      Thin 15              Intel  Core i5 12th Gen 12450H

2           HP       Laptop                AMD  Ryzen 3 Quad Core 5300U

3         Acer          One              Intel  Core i3 11th Gen 1115G4

4           HP           HP                AMD  Ryzen 5 Hexa Core 5500U

5      Infinix      GT Book              Intel  Core i5 12th Gen 12450H

6         Acer     Aspire 7              Intel  Core i5 12th Gen 12450H

7         ASUS  Vivobook 15              Intel   Core i3 12th Gen 1215U

8         Acer     Aspire 3              Intel  Celeron Dual Core N4500

9          MSI    Modern 14                AMD  Ryzen 5 Hexa Core 7530U


  Operating_System  RAM  Storage Storage_Type  Screen_Size  Rating  \
0       Windows 11   16      512          SSD           12     4.4
1       Windows 11   16      512          SSD           12     4.3
2       Windows 11    8      512          SSD           11     4.3
```

```
3       Windows 11    8      512          SSD            11     4.2
4       Windows 11   16      512          SSD            16     4.3
5       Windows 11   16      512          SSD            12     4.4
6       Windows 11   16      512          SSD            12     4.1
7       Windows 11    8      512          SSD            12     4.2
8       Windows 11    8      512          SSD            11     3.8
9       Windows 11   16      512          SSD            16     4.3

   Number_of_Reviews    Price
0               38.0  58990.0
1               34.0  57990.0
2              482.0  30999.0
3              571.0  26990.0
4              268.0  42990.0
5               13.0  64990.0
6              214.0  52990.0
7              360.0  35990.0
8               25.0  20990.0
9              246.0  36990.0
```

```
# Handling the values of Storage column
laptop_df[["Storage"]].describe()
```

```
           Storage
count  1624.000000
mean    440.557266
std     171.792497
min       1.000000
25%     512.000000
50%     512.000000
75%     512.000000
max     512.000000
```

Observation

- The min value of a laptop storage can be 128 GB nothing less than that, if it is less than it means the value is TB and needs to be converted into GB or its a wrong data point.
- Lets filter out all the datapoints where the storage is less than 128 GB.

```
# Filtering out the datapoints where the storage is less than 128
filtered_df = laptop_df[laptop_df['Storage'] < 128]
filtered_df
```

```
     Laptop_Brand                  Laptop_Name Processor_Company  \
17           ASUS            ROG Strix Scar 16             Intel
18             HP   Chromebook MediaTek MT8183       Chromebook
20           Acer                 Predator Neo             Intel
29            MSI                   Claw AI PC             Intel
41           Acer  Acer Predator Helios Neo 16             Intel
...           ...                          ...               ...
1601       Lenovo        Yoga Slim 7x Qualcomm       Snapdragon
```

```
1602           HP   Chromebook MediaTek MT8183        Chromebook
1613       Lenovo                  Yoga AI PC             Intel
1625       Lenovo        Yoga Slim 7x Qualcomm        Snapdragon
1626           HP   Chromebook MediaTek MT8183        Chromebook

                     Processor Operating_System  RAM  Storage
Storage_Type  \
17    Core i9 14th Gen 14900HX       Windows 11   32        2
SSD
18             MediaTek MT8183           Chrome    4       32
EMMC
20    Core i7 13th Gen 13700HX       Windows 11   16        1
SSD
29           Core Ultra 7 155H       Windows 11   16        1
SSD
41    Core i9 13th Gen 13900HX       Windows 11   16        1
SSD
...                         ...              ...  ...      ...
...
1601                   X Elite       Windows 11   32        1
SSD
1602           MediaTek MT8183           Chrome    4       32
EMMC
1613         Core Ultra 7 155H       Windows 11   32        1
SSD
1625                   X Elite       Windows 11   32        1
SSD
1626           MediaTek MT8183           Chrome    4       32
EMMC

      Screen_Size  Rating  Number_of_Reviews     Price
17             14     NaN                NaN  339990.0
18             81     3.8              501.0   11990.0
20             13     4.4               89.0  104990.0
29             16     5.0                1.0   74990.0
41             13     4.2                8.0  134990.0
...           ...     ...                ...       ...
1601           32     NaN                NaN  149990.0
1602           81     3.8              501.0   11990.0
1613           32     NaN                NaN  244890.0
1625           32     NaN                NaN  149990.0
1626           81     3.8              501.0   11990.0

[223 rows x 12 columns]
```

Observation

- EMMC Storage are exception to the traditional laptops as they aare made for extremely light weight load so whatever storage is provided need not to be changed.
- However the storage value in for SSDs/HDDs needs to be updated.

```python
# Converting the Storage values of Storage in TB to GB
condition = (laptop_df['Storage_Type'].isin(['SSD', 'HDD'])) &
(laptop_df['Storage'] < 128)
laptop_df.loc[condition, 'Storage'] *= 1024

# Looking at Data Sumamry
laptop_df.describe(include="all").T
```

```
                    count unique                       top  freq
mean  \
Laptop_Brand         1624      9                        HP   388
NaN
Laptop_Name          1624     47                        HP   136
NaN
Processor_Company    1624      4                     Intel  1032
NaN
Processor            1624     38  Core i3 12th Gen 1215U   182
NaN
Operating_System     1624      2              Windows 11  1533
NaN
RAM                1624.0    NaN                       NaN   NaN
12.23399
Storage            1624.0    NaN                       NaN   NaN
526.857143
Storage_Type         1624      2                       SSD  1533
NaN
Screen_Size        1624.0    NaN                       NaN   NaN
19.899631
Rating             1495.0    NaN                       NaN   NaN
4.203144
Number_of_Reviews  1495.0    NaN                       NaN   NaN
247.48495
Price              1624.0    NaN                       NaN   NaN
53425.598522

                           std      min      25%      50%      75%
max
Laptop_Brand               NaN      NaN      NaN      NaN      NaN
NaN
Laptop_Name                NaN      NaN      NaN      NaN      NaN
NaN
Processor_Company          NaN      NaN      NaN      NaN      NaN
NaN
Processor                  NaN      NaN      NaN      NaN      NaN
NaN
Operating_System           NaN      NaN      NaN      NaN      NaN
NaN
RAM                   5.873543      4.0      8.0      8.0     16.0
32.0
Storage             202.351364     32.0    512.0    512.0    512.0
```

```
2048.0
Storage_Type                NaN       NaN       NaN       NaN       NaN
NaN
Screen_Size           18.594559     11.0      11.0      12.0      16.0
81.0
Rating                 0.239443      3.3       4.1       4.2       4.3
5.0
Number_of_Reviews    202.948047      0.0      25.0     214.0     467.0
597.0
Price               49743.83399  11990.0  30999.0  37999.0  54990.0
339990.0
```

Observation

- Based on the above describe(Mean, precentiles and median) of the storage column, we can be sure that all the values are now accurate.

```
# Dealing with Screen Size outlier values
laptop_df[laptop_df["Screen_Size"] > 17].head()
```

```
   Laptop_Brand                 Laptop_Name Processor_Company  \
12           HP                         15s               AMD
18           HP  Chromebook MediaTek MT8183        Chromebook
36           HP                         15s               AMD
42           HP  Chromebook MediaTek MT8183        Chromebook
60           HP                         15s               AMD

                    Processor Operating_System  RAM  Storage
Storage_Type  \
12  Ryzen 3 Quad Core 5300U       Windows 11    8      512
SSD
18          MediaTek MT8183           Chrome    4       32
EMMC
36  Ryzen 3 Quad Core 5300U       Windows 11    8      512
SSD
42          MediaTek MT8183           Chrome    4       32
EMMC
60  Ryzen 3 Quad Core 5300U       Windows 11    8      512
SSD

    Screen_Size  Rating  Number_of_Reviews    Price
12           64     4.2              405.0  32490.0
18           81     3.8              501.0  11990.0
36           64     4.2              405.0  32490.0
42           81     3.8              501.0  11990.0
60           64     4.2              405.0  32490.0
```

```
laptop_df[laptop_df["Screen_Size"] > 17].shape
```

```
(265, 12)
```

Observation

- The Screen Size values have been mislabelled during the aggregation or were not present in the give data.
- In order to impute these outlier values, these values will be replaced by the mean value of screen size based on each laptop company

```
# Filtering out what all values are there in Screen Size
screen_size_counts =
laptop_df['Screen_Size'].value_counts().sort_index()

screen_size_counts
```

```
Screen_Size
11    485
12    403
13    218
14      5
16    248
25     22
32     62
64    113
81     68
Name: count, dtype: int64
```

Observation

- The above values confirm that some values are mis-represented, so we will be replacing it with the mean value based on each laptop.
- This will be done for screen size greater than 32, and for values between 17 and 32 will be converted to inches.

```
# Imputing outliers in Screen_Size Column

# Function to replace screen sizes greater than 32 with the median
screen size for each brand
# and convert screen sizes between 17 and 32 from cm to inches
def update_screen_size(group):
    median_size = group['Screen_Size'].median()
    group.loc[group['Screen_Size'] > 32, 'Screen_Size'] = median_size
    group.loc[(group['Screen_Size'] > 17) & (group['Screen_Size'] <=
32), 'Screen_Size'] *= 0.393701
    return group

# Apply the function to each group of Laptop_Brand
laptop_df =
laptop_df.groupby('Laptop_Brand').apply(update_screen_size).reset_inde
x(drop=True)

# Filtering out what all values are there in Screen Size
screen_size_counts =
```

```
laptop_df['Screen_Size'].value_counts().sort_index()
screen_size_counts
```

```
Screen_Size
9.842525      22
11.000000    507
12.000000    403
12.598432     62
13.000000    218
14.000000      5
16.000000    407
Name: count, dtype: int64
```

Observation

- Need to replace all the screen size values less than 11 to 11 inches
- Make the the screen size to definitive 12.5 inches

```
# Replace specific Screen_Size values
laptop_df['Screen_Size'] = laptop_df['Screen_Size'].replace({9.842525:
11.00, 12.598432: 12.5})

# Filtering out what all values are there in Screen Size
screen_size_counts =
laptop_df['Screen_Size'].value_counts().sort_index()
screen_size_counts
```

```
Screen_Size
11.0    529
12.0    403
12.5     62
13.0    218
14.0      5
16.0    407
Name: count, dtype: int64
```

Observation

- All the screen sizes are now valid.

```
laptop_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1624 entries, 0 to 1623
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Laptop_Brand       1624 non-null   object
 1   Laptop_Name        1624 non-null   object
 2   Processor_Company  1624 non-null   object
 3   Processor          1624 non-null   object
```

```
 4   Operating_System   1624 non-null   object
 5   RAM                1624 non-null   int64
 6   Storage            1624 non-null   int64
 7   Storage_Type       1624 non-null   object
 8   Screen_Size        1624 non-null   float64
 9   Rating             1495 non-null   float64
 10  Number_of_Reviews  1495 non-null   float64
 11  Price              1624 non-null   float64
dtypes: float64(4), int64(2), object(6)
memory usage: 152.4+ KB
```

```
laptop_df.describe(include="all").T
```

```
                     count unique                      top  freq
mean  \
Laptop_Brand          1624      9                       HP   388
NaN
Laptop_Name           1624     47                 Aspire 7   136
NaN
Processor_Company     1624      4                    Intel  1032
NaN
Processor             1624     38  Core i3 12th Gen 1215U   182
NaN
Operating_System      1624      2              Windows 11  1533
NaN
RAM                 1624.0    NaN                     NaN   NaN
12.23399
Storage             1624.0    NaN                     NaN   NaN
526.857143
Storage_Type          1624      2                     SSD  1533
NaN
Screen_Size         1624.0    NaN                     NaN   NaN
12.836207
Rating              1495.0    NaN                     NaN   NaN
4.203144
Number_of_Reviews   1495.0    NaN                     NaN   NaN
247.48495
Price               1624.0    NaN                     NaN   NaN
53425.598522

                           std      min      25%      50%      75%
max
Laptop_Brand               NaN      NaN      NaN      NaN      NaN
NaN
Laptop_Name                NaN      NaN      NaN      NaN      NaN
NaN
Processor_Company          NaN      NaN      NaN      NaN      NaN
NaN
Processor                  NaN      NaN      NaN      NaN      NaN
NaN
```

```
Operating_System           NaN       NaN       NaN       NaN       NaN
NaN
RAM                   5.873543       4.0       8.0       8.0      16.0
32.0
Storage             202.351364      32.0     512.0     512.0     512.0
2048.0
Storage_Type               NaN       NaN       NaN       NaN       NaN
NaN
Screen_Size            1.94802      11.0      11.0      12.0      16.0
16.0
Rating                0.239443       3.3       4.1       4.2       4.3
5.0
Number_of_Reviews   202.948047       0.0      25.0     214.0     467.0
597.0
Price              49743.83399   11990.0   30999.0   37999.0   54990.0
339990.0
```

Observation

- Replacing all the missing values in `Rating` and `Number_of_Reviews` with zero, as they will be treated as the laptop not being sold or not that appealing to customers.

```
# Replacing Nan values in the Rating and Number of Reviews column
laptop_df['Rating'] = laptop_df['Rating'].fillna(0)
laptop_df['Number_of_Reviews'] =
laptop_df['Number_of_Reviews'].fillna(0)

# Taking a final look at info and description of data
laptop_df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1624 entries, 0 to 1623
Data columns (total 12 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Laptop_Brand       1624 non-null   object
 1   Laptop_Name        1624 non-null   object
 2   Processor_Company  1624 non-null   object
 3   Processor          1624 non-null   object
 4   Operating_System   1624 non-null   object
 5   RAM                1624 non-null   int64
 6   Storage            1624 non-null   int64
 7   Storage_Type       1624 non-null   object
 8   Screen_Size        1624 non-null   float64
 9   Rating             1624 non-null   float64
 10  Number_of_Reviews  1624 non-null   float64
 11  Price              1624 non-null   float64
dtypes: float64(4), int64(2), object(6)
memory usage: 152.4+ KB

laptop_df.describe(include="all").T
```

	count	unique	top	freq	mean
Laptop_Brand	1624	9	HP	388	NaN
Laptop_Name	1624	47	Aspire 7	136	NaN
Processor_Company	1624	4	Intel	1032	NaN
Processor	1624	38	Core i3 12th Gen 1215U	182	NaN
Operating_System	1624	2	Windows 11	1533	NaN
RAM	1624.0	NaN	NaN	NaN	12.23399
Storage	1624.0	NaN	NaN	NaN	526.857143
Storage_Type	1624	2	SSD	1533	NaN
Screen_Size	1624.0	NaN	NaN	NaN	12.836207
Rating	1624.0	NaN	NaN	NaN	3.869273
Number_of_Reviews	1624.0	NaN	NaN	NaN	227.826355
Price	1624.0	NaN	NaN	NaN	53425.598522

	std	min	25%	50%	75%	max
Laptop_Brand	NaN	NaN	NaN	NaN	NaN	NaN
Laptop_Name	NaN	NaN	NaN	NaN	NaN	NaN
Processor_Company	NaN	NaN	NaN	NaN	NaN	NaN
Processor	NaN	NaN	NaN	NaN	NaN	NaN
Operating_System	NaN	NaN	NaN	NaN	NaN	NaN
RAM	5.873543	4.0	8.0	8.0	16.0	32.0
Storage	202.351364	32.0	512.0	512.0	512.0	2048.0
Storage_Type	NaN	NaN	NaN	NaN	NaN	NaN
Screen_Size	1.94802	11.0	11.0	12.0	16.0	16.0
Rating	1.159917	0.0	4.075	4.2	4.3	5.0
Number_of_Reviews	205.902161	0.0	12.0	214.0	424.0	

```
597.0
Price               49743.83399  11990.0  30999.0  37999.0  54990.0
339990.0

# Since the data is now cleaned, it can be exported as a clean CSV for
further analysis
laptop_df.to_csv(r"data\flipkart_laptop_cleaned.csv",index=False)
```

Data Collection and Cleaning - Best Buy

- The dataset was scrapped from Best Buy website.
- Scraping and Cleaning done by:
  - Name: Yeswanth Chitturi
  - UB ID: 50591666

1. Importing Modules

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
```

2. Scraping All The Webpages Of Best Buy For Laptop Data

Scraping Code , Converting Html to data frame and saving as CSV.

```
# Function to get the HTML content of a page
def get_page_content(url):
    headers = {'User-Agent': 'Mozilla/5.0'}
    response = requests.get(url, headers=headers)
    return BeautifulSoup(response.content, 'html.parser')

# Function to scrape laptop data from one page
def scrape_laptop_data_from_page(soup, data):
    base_url = 'https://www.bestbuy.com'

    # Find all laptops on the page
    laptops = soup.find_all('li', class_='sku-item')

    for laptop in laptops:
        try:
            # Extract laptop name
            name_tag = laptop.find('h4', class_='sku-title')
            name = name_tag.text.strip() if name_tag else 'No name
available'

            # Extract SKU value
```

```
            skuvalue_tag = laptop.find('span', class_='sku-value')
            skuvalue = skuvalue_tag.text.strip() if skuvalue_tag else
'No SKU value available'

            # Extract rating (visually hidden text)
            visually_hidden_tag = laptop.find('p', class_='visually-
hidden')
            rating = visually_hidden_tag.text.strip() if
visually_hidden_tag else 'No rating available'

            # Extract laptop price
            price_tag = laptop.find('div', class_='priceView-customer-
price')
            price = price_tag.span.text.strip() if price_tag else 'No
price available'

            # Append data to the list
            data.append([name, skuvalue, rating, price])
            #print(f'Scraped: {name}')

        except Exception as e:
            print(f'Error scraping laptop: {e}')

# Function to find the link to the next page
def get_next_page(soup):
    next_page_tag = soup.find('a', class_='sku-list-page-next')
    if next_page_tag and 'href' in next_page_tag.attrs:
        return next_page_tag['href']
    return None

# Main function to scrape all pages
def scrape_all_pages():
    base_url = 'https://www.bestbuy.com'
    search_url = f'{base_url}/site/searchpage.jsp?st=laptops'

    # List to hold all laptop data
    data = []

    current_page_url = search_url

    while current_page_url:
        #print(f'Scraping page: {current_page_url}')
        soup = get_page_content(current_page_url)
        scrape_laptop_data_from_page(soup, data)

        # Check if there's a next page
        next_page = get_next_page(soup)
        if next_page:
            current_page_url = f'{base_url}{next_page}'
        else:
```

```
            current_page_url = None

    # Create a DataFrame from the scraped data
    df = pd.DataFrame(data, columns=['total_info', 'Model No',
'Rating', 'Price'])

    # Save DataFrame to a CSV file
    df.to_csv('laptops_data.csv', index=False)
    print('Data has been saved to laptops_data.csv')
    return df

if __name__ == '__main__':
    df=scrape_all_pages()
    #print(df)
print(df.info())
```

```
Data has been saved to laptops_data.csv
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1285 entries, 0 to 1284
Data columns (total 4 columns):
 #   Column      Non-Null Count  Dtype
---  ------      --------------  -----
 0   total_info  1285 non-null   object
 1   Model No    1285 non-null   object
 2   Rating      1285 non-null   object
 3   Price       1285 non-null   object
dtypes: object(4)
memory usage: 40.3+ KB
None
```

Data Collection Similar to Flipkart data

Converting Raw Data Into Columns

```
#Extracting Columns
import pandas as pd
#Brand
df['Brand']='-'
df['Brand'] = df['total_info'].str.split(' ').str[0]
#Colur
df['Colour']='-'
df['Colour'] = df['total_info'].str.split(' ').str[-1]
#Saving original
df1=df
#Remove Special Characters
df1['total_info'] = df1['total_info'].str.replace('”', '"',
regex=False)
#Ram
ram_pattern = r'(\s*\d+\s*G[Bb]\s*(Memory|RAM|Ram)\s*|(\d{1,2}GB))'
```

```
df1['ram'] = df1['total_info'].str.extract(ram_pattern, expand=False)
[0]
df1['ram'] = df1['ram'].str.extract(r'(\d+)').astype(float)
df1['ram'] = df1['ram'].apply(lambda x: f"{int(x)} GB" if x < 33 else
pd.NA)
#Storage
storage_pattern = r'(\d+\s*(TB|GB|G)\s*(SSD|HDD|Solid State Drive|
Flash Storage|Hard Drive|eMMC|UFS|SDD|PCIe|NVMe|Storage)|\d+\s*TB\s*-\
s*SSD|\d{3,4}SSD|\d{3,4}GB\s*-\s*SSD)'
df1['storage'] = df1['total_info'].str.extract(storage_pattern,
expand=False)[0]
#Processor
processor_pattern = r'(Intel\s+\w+\s*\w*|\bM[12]\s+(?:Pro|Max|chip)\
s*Built\s*for\s*Apple\s*-\w+|M[12]\s+(?:Pro|Max|chip)\b|M3\s+chip\
s*Built\s*for\s*Apple\s*Intelligence|M3\s+\w+\s+chip\s*Built\s*for\
s*Apple|M3\s+chip|AMD\s+Ryzen\s+\d+(?:-\d+|\s+\d+)\w*|AMD\s+Ryzen\s+\
d+\b|AMD\s+Ryzen\s+AI\s+\d+-\d+\s+\w*|Apple\s+M1\s+\w+\s+chip\b|Core\
s+\w+\s*-\d+|\bRyzen\s+\d+\s+\d+\w*|MediaTek\s+\w+\s+\d+|Snapdragon\
s+\w+\s*\w*|Pentium\s+\w+\s+\d+)'
extracted_processors =
df1['total_info'].str.extract(processor_pattern)
df1['processor'] = extracted_processors[0].fillna(value=pd.NA)
#Display size
display_pattern=r'(\d{2}"|\d{2}.\d"|\d{2}-inch|\d{2}-Inch|\d{2}.\d-
inch|\d{2}.\d-Inch)'
extracted_displays = df1['total_info'].str.extract(display_pattern)
df1['display'] = extracted_displays[0].fillna(value=pd.NA)
#Laptop Name
model_pattern = r'(Geek Squad Certified Refurbished MacBook Air|Envy|
XPS|OmniBook|ProBook|Flex|LOQ|Katana|Blade|Aero|Vector|Summit|Raider|
Pulse|Cyborg|Elitebook|Precision|Galaxy book|Geek Squad Certified
Refurbished Macbook®|Geek Squad Certified Refurbished MacBook Pro|
MacBook Pro|GSRF MacBook Pro|MacBook Air|Refurbished MacBook®|
Chromebook|ProArt P16|ProArt Studiobook|ROG Strix G16|ROG Strix SCAR|
ROG Zephyrus G14|TUF A15|TUF Gaming A17|Vivobook|Vivobook Pro 15|
Zenbook|Zenbook Pro Duo|Zenbook DUO Dual|Nitro 16|Predator Helios 18|
Predator Triton|Swift X|TravelMate|m18 R2|m16|m18 165|Latitude 7000|
Refurbished|Inspiron|Precision 3540|XPS 13|Latitude 3000|Mobile
Processor|Alienware x14|Gaming Laptop IPS|EliteBook 840 G8|Elite x360
830 G11|Envy 2-in-1|Pavilion|Victus|ZBook|OMEN|gram|gram SuperSlim|
ThinkPad X1|Yoga|Yoga Book|Yoga Pro|Legion|Ideapad|ThinkPad T14s|GSRF
Surface|Surface Laptop - Copilot\+ PC|Surface Laptop Studio 2|Bravo
15|Commercial 14 H A13MG|Creator 17|Crosshair|Raider GE78 HX|Stealth|
Summit E16 AI Studio|Vector 16 HX A14V|Blade 16|Galaxy Book2 Pro|
Galaxy Book3 360|Galaxy Book4 Ultra|Geek Squad Certi Refurbis Galaxy
Book Flex2 Alpha|Galaxy Book4 Edge|NEOZ3 Laptop|Galaxy Book4 Pro 360|
Blade 18|ROG Zephyrus|Summit E16 AI Evo|Raider GE68 HX|Prestige|Modern
15|ThinkPad|ThinkPad T14|ThinkBook 14 G7|Spectre|AORUS|Latitude 5000|
Latitude|Predator Helios|Aspire 5|ProArt PX13 13|Blade 15|ProBook 445|
```

```
EliteBook 640|Raider 18|CreatorPro|Creator 16|ThinkBook|Swift 14|ROG
Flow)'
df1['model'] = df1['total_info'].str.extract(model_pattern)
#Graphics
graphics_pattern = r'(NVIDIA GeForce RTX \d+|Ryzen \d+U|Intel Iris Xe
Graphics|NVIDIA Quadro P1000|Intel UHD Graphics 620|NVIDIA Quadro
T1000|AMD Radeon Graphics|NVIDIA GeForce RTX4070|NVIDIA Quadro P3200)'
df1['graphics'] = df1['total_info'].str.extract(graphics_pattern)
#Storage Type
storagetype_pattern = r'(FlashStorage|HardDrive|SSD|HDD|SDD|eMMC|Flash
Storage|Hard Drive|NVMe|PCIe|UFS|Solid State Drive)'
df1['storage_type'] = df1['storage'].str.extract(storagetype_pattern)
#No of reviews
no_reviews_pattern = r'(\d{1,4})\s*reviews?|reviewfalse'
df1['no_reviews'] = df1['Rating'].str.extract(no_reviews_pattern)
#Rating out of 5
Rating_5_pattern = r'Rating\s*(\d(?:\.\d)?)'
df1['Rating_5'] = df1['Rating'].str.extract(Rating_5_pattern)
#Processor Company
processor_company_pattern = r'(Intel|AMD\s*Ryzen|Ryzen|Snapdragon|
Core|MediaTek|M1|M2|M3)'
df1['processor_company'] =
df1['processor'].str.extract(processor_company_pattern)
#Operating System
df1['os']='Windows11'
#Dropping group columns
df1 = df1.drop('total_info', axis=1)
df1 = df1.drop('Rating', axis=1)
#Saving original
df_BB=df1
```

Filling Null Values

```
df_BB['graphics'] = df_BB['graphics'].fillna('No Graphics')

df_BB['no_reviews'] = df_BB['no_reviews'].fillna('0')

df_BB['Rating_5'] = df_BB['Rating_5'].fillna('0 Reviews')

df_BB['storage_type'] = df_BB['storage_type'].fillna('SSD')

df_BB['processor'] = df_BB['processor'].fillna('No Info')

df_BB['processor_company'] = df_BB['processor_company'].fillna('No
Info')

df_BB['model'] = df_BB['model'].fillna('No Model')

most_frequent_display = df_BB['display'].mode()[0]
df_BB['display'] = df_BB['display'].fillna(most_frequent_display)
```

```
most_frequent_ram = df_BB['ram'].mode()[0]
df_BB['ram'] = df_BB['ram'].fillna(most_frequent_ram)

most_frequent_storage = df_BB['storage'].mode()[0]
df_BB['storage'] = df_BB['storage'].fillna(most_frequent_storage)
```

Removing extra names and characters

```
#More cleaning

df_BB.loc[df_BB['Brand'] == 'Apple', 'os'] = 'Mac OS'

df_BB['processor_company'] = df_BB['processor_company'].replace({'M1':
'Apple', 'M2': 'Apple', 'M3': 'Apple','Core':'Intel','Ryzen': 'AMD
Ryzen'})

df_BB['processor'] = df_BB['processor'].str.replace('Intel', '')
df_BB['processor'] = df_BB['processor'].str.replace('AMD Ryzen', '')
df_BB['processor'] = df_BB['processor'].str.replace('Snapdragon', '')
df_BB['processor'] = df_BB['processor'].str.replace('MediaTek', '')
df_BB['processor'] = df_BB['processor'].str.replace('Ryzen', '')

df_BB['Brand'] = df_BB['Brand'].str.replace('New!', '')

df_BB['display'] = df_BB['display'].str.replace('"', '')

df_BB['storage'] = df_BB['storage'].str.replace('GB', '')
df_BB['storage'] = df_BB['storage'].str.replace('SSD', '')
df_BB['storage'] = df_BB['storage'].str.replace('Solid State Drive',
'')
df_BB['storage'] = df_BB['storage'].str.replace('SDD', '')
df_BB['storage'] = df_BB['storage'].str.replace('HDD', '')
df_BB['storage'] = df_BB['storage'].str.replace('TB PCIe', 'TB')
df_BB['storage'] = df_BB['storage'].str.replace('-', '')
df_BB['storage'] = df_BB['storage'].str.replace(' ', '')
df_BB['storage'] = df_BB['storage'].str.replace('HardDrive', '')
df_BB['storage'] = df_BB['storage'].str.replace('NVMe', '')
df_BB['storage'] = df_BB['storage'].str.replace('FlashStorage', '')
df_BB['storage'] = df_BB['storage'].str.replace('TB NVMe', 'TB')
df_BB['storage'] = df_BB['storage'].str.replace('Storage', '')
def convert_tb_to_gb(storage_value):
    if 'TB' in storage_value:

        tb_value = float(storage_value.replace('TB', '').strip())
        return str(int(tb_value * 1024))
    return storage_value
df_BB['storage'] = df_BB['storage'].apply(convert_tb_to_gb)

df_BB['storage_type'] = df_BB['storage_type'].str.replace('Solid State
```

```
Drive', 'SSD')

df_BB['ram'] = df_BB['ram'].str.replace('GB', '')
df_BB['ram'] = df_BB['ram'].str.replace(' ', '')

#Merging Model columns
df_BB['processor_model'] = df_BB['processor'] + ' - ' + df_BB['Model
No']
df_BB = df_BB.drop(['processor', 'Model No'], axis=1)

df_BB['Price'] = df_BB['Price'].str.replace('$', '', regex=False)

df_BB['Price'] = df_BB['Price'].str.replace(',', '')

df_BB['display'] = df_BB['display'].str.replace('-Inch', '')

df_BB.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1285 entries, 0 to 1284
Data columns (total 14 columns):
 #   Column             Non-Null Count  Dtype
---  ------             --------------  -----
 0   Price              1285 non-null   object
 1   Brand              1285 non-null   object
 2   Colour             1285 non-null   object
 3   ram                1285 non-null   object
 4   storage            1285 non-null   object
 5   display            1285 non-null   object
 6   model              1285 non-null   object
 7   graphics           1285 non-null   object
 8   storage_type       1285 non-null   object
 9   no_reviews         1285 non-null   object
 10  Rating_5           1285 non-null   object
 11  processor_company  1285 non-null   object
 12  os                 1285 non-null   object
 13  processor_model    1285 non-null   object
dtypes: object(14)
memory usage: 140.7+ KB
```

Renaming as required

```
#Renameing columns for better understading
df_BB = df_BB.rename(columns={'model': 'Laptop_name'})
df_BB = df_BB.rename(columns={'Price': 'Laptop_Price'})
df_BB = df_BB.rename(columns={'Brand': 'Laptop_Brand'})
df_BB = df_BB.rename(columns={'Price': 'Laptop_Price'})
df_BB = df_BB.rename(columns={'Colour': 'Laptop_Colour'})
df_BB = df_BB.rename(columns={'ram': 'Laptop_Memory_GB'})
df_BB = df_BB.rename(columns={'storage': 'Laptop_Storage_GB'})
```

```
df_BB = df_BB.rename(columns={'display': 'Laptop_Display_Size_in'})
df_BB = df_BB.rename(columns={'graphics': 'Laptop_Graphics'})
df_BB = df_BB.rename(columns={'no_reviews': 'No_Of_Reviews'})
df_BB = df_BB.rename(columns={'processor_company':
'Processor_Company_Name'})
df_BB = df_BB.rename(columns={'storage_type': 'Storage_Type'})
df_BB = df_BB.rename(columns={'os': 'Operating_System'})
df_BB = df_BB.rename(columns={'processor_model': 'Processor_Model'})

df_BB.head()
```

```
  Laptop_Price Laptop_Brand Laptop_Colour Laptop_Memory_GB
Laptop_Storage_GB  \
0       399.99         Dell         Black                8
512
1       549.99         Dell         Black               16
1024
2       799.99           HP        Silver               16
512
3      1099.99           HP        Silver               16
1024
4       329.99       Lenovo          Blue                8
256

  Laptop_Display_Size_in Laptop_name      Laptop_Graphics Storage_Type
\
0                     14    Inspiron          No Graphics          SSD

1                     14    Inspiron          No Graphics          SSD

2                   15.6    No Model          No Graphics          SSD

3                     14        Envy          No Graphics          SSD

4                   15.6     Ideapad  AMD Radeon Graphics          SSD


  No_Of_Reviews Rating_5 Processor_Company_Name Operating_System  \
0           646      4.5                  Intel        Windows11
1            86      4.7                  Intel        Windows11
2          1962      4.6                  Intel        Windows11
3           282      4.8                  Intel        Windows11
4           193      4.5              AMD Ryzen        Windows11

                Processor_Model
0   Core i5 - i3520-5124BLK-PUS
1   Core i7 - i3520-7896BLK-PUS
2         Core i7 - 15-DY5073DX
3      Core Ultra - 14-fc0023dx
4          5 7520U - 82VG00MYUS
```

Explanation for Modules

- `requests` module is to send a request to the URL to fetch the data.
- `BeautifulSoup` is a class using the object of which we will deal with the scraped HTML data.
- `re` is for using regex patterns to filter out data and create our dataframe in an organized format.
- `numpy` and `pandas` module if for manipulating data values and handling the data overall.

Data Collection and Cleaning - Amazon

- The dataset was scrapped from the Amazon website.
- Scraping and Cleaning done by:
  - Name: Shaurya Mathur
  - UB ID: 50611201

Data Collection

Scraping the URLs from Amazon website across 70 pages and fetch URLs from the paginate response

All URLs will be saved to a file and will be parsed one-by-one for detailed product specs.

Amazon Website example paginated page -

https://www.amazon.com/s?i=computers&rh=n%3A565108&fs=true&page=3&qid=1726990472

NOTE - The following data collection scripts were prepared and executed outside this python notebook in my local IDE. The approximate execution time was approximately 5 hours and was run in batches. The cells with data collection scripts have not been executed in this notebook.

```
import requests
from bs4 import BeautifulSoup
import time
import random

headers = {
    'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/91.0.4472.124
Safari/537.36',
    'Accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/
*;q=0.8',
    'Accept-Language': 'en-US,en;q=0.5',
```

```
    'Accept-Encoding': 'gzip, deflate, br',
    'Connection': 'keep-alive',
    'Upgrade-Insecure-Requests': '1',
    'Cache-Control': 'max-age=0'
}
all_laptop_urls = set()

# Function to scrape a single page
def scrape_page(url, existing_urls):

    response = requests.get(url, headers=headers)

    if response.status_code == 200:
        soup = BeautifulSoup(response.content, 'html.parser')
        product_links = soup.find_all('a', class_='a-link-normal s-
underline-text s-underline-link-text s-link-style a-text-normal')
        laptop_urls = [link.get('href') for link in product_links if
link.get('href')]
        laptop_urls = ['https://www.amazon.com' + url if
url.startswith('/') else url for url in laptop_urls]
        # Filter out URLs that are already in existing_urls
        new_laptop_urls = [url for url in laptop_urls if url not in
existing_urls]

        return new_laptop_urls
    else:
        print(f"Failed to download page {page_number}. Status code:
{response.status_code}")
        return []

# Loop through 70 pages
for page_number in range(70):
    print(f"Scraping page {page_number}...")
    url = f'https://www.amazon.com/s?i=computers&rh=n
%3A565108&fs=true&page={page_number}&qid=1726990472'
    page_urls = scrape_page(url, all_laptop_urls)

    # Check if the page has less than 20 URLs.
    # Sometimes due to ads and different product categories - Amazon
displays products in the range of 20-24 URLs
    # This check is added to ensure we scrape more than 20 URLs per
pagination
    if len(page_urls) < 20:
        print(f"Warning: Page {page_number} has less than 20 URLs.")
        print(f"URL: {url}")
        # Added to manually check what went wrong with the URL, why it
has less than required URLs
        input("Press Enter to continue...")

    # Add URLs to the set
```

```
    all_laptop_urls.update(page_urls)

    # Append URLs to a file - we will iterate and scrape these again
to fetch product specific details.
    with open('laptop_urls.txt', 'a') as file:
        for url in page_urls:
            file.write(url + '\n')

    # Add a delay between requests to avoid overwhelming the server
and getting blocked
    time.sleep(random.uniform(1, 5))

print(f"Scraped a total of {len(all_laptop_urls)} unique laptop
URLs.")
print(f"All URLs have been appended to laptop_urls.txt")
```

Read the .txt file and scrape individual product specs.

```
import requests
from bs4 import BeautifulSoup
import time
import random
import os
import csv

# This was a time consuming script and was thus run in multiple
batches
# Output file name is to ensure each batch gets its own file name, we
will combine these batches using pandas during the cleaning process.
output_file = 'final_laptop_data_unprocessed_4.csv'

# For ensuring correct batching while reading URLs from the file.
start_index = 0
end_index = 100

# Amazon blocks if same user agent is used. To avoid this, I use user
agents in random order to avoid this.
user_agents = [
'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/91.0.4472.124 Safari/537.36',
'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/605.1.15
(KHTML, like Gecko) Version/14.1.1 Safari/605.1.15',
'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like
Gecko) Chrome/92.0.4515.107 Safari/537.36',
'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/129.0.0.0 Safari/537.36'
]

def scrape_laptop_info(url):
```

```
    headers = {
        'User-Agent': random.choice(user_agents),
        'Accept':
'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/
*;q=0.8',
        'Accept-Language': 'en-US,en;q=0.5',
        'Accept-Encoding': 'gzip, deflate, br',
        'Connection': 'keep-alive',
        'Upgrade-Insecure-Requests': '1',
        'Cache-Control': 'max-age=0',
        'Cookie':'x-amz-captcha-1=1724354575099927; x-amz-captcha-
2=SQkSw6jFuY811RLzQyZV3g==; skin=noskin; ubid-main=132-1074986-
8554950; lc-main=en_US; csd-
key=eyJ3YXNtVGVzdGVkIjp0cnVlLCJ3YXNtQ29tcGF0aWJsZSI6dHJ1ZSwid2ViQ3J5cH
RvVGVzdGVkIjpmYWxzZSwidiI6MSwia2lkIjoiMThmMWM5Iiwia2V5IjoiYy9yQnIvVHZW
SjhpNGx6VFJPcTRnbXdnY0o5dzFEUzdmZTk0ekZpY2NPUWtQQUd4ZS8rYXdkMjB1OEN4T1
VTemVQWFRzYnhjL0FNOGo2UU41ekJVbldSV0ZEd1RheXlHZTBJa3d0NHc3YjBia2hpVWVY
T1M5Z2swRGJPWXVSd0t1T3djOTlXVXZUUWVzUWRCWHdWQVBkT3QvUWNpekpmKzZrT3gvL1
d6WmorRHNPSC9uckxYclFQY3h1Rm5lRlIwOXpONlBVT1dGdXRmVHR5L2djVzJWT2N2dngr
amdvUDlTdElJL1AvM0JvZWJSUWQ4WFNncGZkTWlWUDFSSjdpU1NSZlNlS1NvQUY3ZHhHSk
NKMEFCUnVGVTNSUG4ycERSWC9DWDYxVlJ2RktwNUEyV3NmUW8wdkM2ZzRwVjZRMHc4OTJT
NjhPaDlqV05kRzFENWEzYkZnPT0ifQ==; lwa-
context=ac5bd3cbe61418b1832e49014f3e7c05; i18n-prefs=USD; sst-
main=Sst1|
PQE0QkvG34FUSgYtJgJArAqFCQuDYhebhfYUa6kPdNW0qF73UEfzaoj7vWo70661JUj9iT
3NVJnLAGey1Vu62ELv8jhW-
Xu0db_RoAYZygvgnTEhBva0BXoHPUE5T_6VaO8u5NZl7h0GhcS1CtEWI9TbdGRtJo3mydk
faSlOjytRkyjG8fNAY7VKlw-5V-
P2QshWpQIni6GwbubRdEDioT1wqIuny26Wx9Epv8qUe7csttn_Tj7ZRMKWYcZ-
5iqvd2UE_aREMysZ8yw89GOPtDNNkoCn9C0h91sc0xvW2DlbXD4; x-
main=1Bj0FFwf0izbYhJyvygPfiMsNhZq1WMa9YhUbAGsMfrLnJLqkJiWWM7NeG38OT0c;
at-main=Atza|IwEBIIYyclilis5nOa1AG_ldOLl-
m2oFeOJQeaPf9NXc_HaGOWx72F9jH71zdnEKa53MF354Z_1ILc8g17RazrFy_ZzEZ1nYnH
MESsi1P19cYkGUgcNcOcyDUdErm1OBVrcVTaRF5DY-
b_7VvVinFkcRc0vTYktj7lBjOF7TgM4pzX03kCSElKAtBYHLk6EObfXy8OqJqcrfIxjys_
qtYn93XGDDQJRpzotBEyjb0el5nKU54qRurM3elbvp92SkUoWDMrM; sess-at-
main="n+5GZmKXcpknPmvMQpYmWwmzzzjYJMs3LP2f6wbzMkU="; session-id=136-
6041417-7446714; session-id-apay=136-6041417-7446714; session-id-
time=2082787201l; JSESSIONID=99F170CEE959AE4A7A90B363B862ECDA;
session-token=WxnfLybBYAHQ8FqN6VBffFF58LA2ClevHEfK1kW9Rka0z/HbRKtTjXi1
o3kLQqAV7dJixuYExxpm6nIFMBUo33bAv23Zjt8yD95rLKmyTmYMS8hIQFCvcZypkzDFsp
ut8AFpRnP5YQ074K+R7wq8zEUzfIHxA5dK3kpUx2TeDxD7UNRTOaod5hF30DeTj8ctwT/
dKGNDkIyeEcbRSfpB1Zrs26T938YkFd0JiFk4jvk4EXSJvMOblzDPQ9wgLv712Rd/
hJYwyxBwvt4L6ArJ6elH2AYWf8vx8EuqAyKOCVIssXOPIPWDjxenCACggACK1Y5Wj/
LOm3OgaLu2/rQd/qrocGHvZ9ac41f8ZedQ4+9T0getVKPU2tRKptdP/Zm+; csm-
hit=tb:s-B605WMKRTQ0E45CQ0KMS|
1727042773231&t:1727042775886&adb:adblk_yes'
    }
```

```
    response = requests.get(url, headers=headers)
    if response.status_code != 200:
        print(f"Failed to fetch the page. Status code:
{response.status_code}")
        return

    soup = BeautifulSoup(response.content, 'html.parser')
    # Extract product information
    title = soup.find('span', {'id': 'productTitle'}).text.strip() if
soup.find('span', {'id': 'productTitle'}) else 'N/A'
    if title == 'N/A':
        return None
    try:
        priceWhole = soup.find('span', class_='a-price-
whole').text.replace(',', '').strip()
        priceDecimal = soup.find('span', class_='a-price-
fraction').text.strip()

        if priceWhole != 'N/A' and priceDecimal != 'N/A':
            price = f"{float(priceWhole) +
float(priceDecimal)/100:.2f}"
        else:
            price = 'N/A'
    except AttributeError:
        price = 'N/A'

    rating = soup.find('span', {'class': 'a-icon-alt'}).text.strip()
if soup.find('span', {'class': 'a-icon-alt'}) else 'N/A'

    typicalPrice = 'N/A'
    # Look for 'Typical Price' or 'List Price' - This is the usual
price without the current deal/discount available.
    price_labels = soup.find_all('span', class_='a-size-small aok-
offscreen')
    for label in price_labels:
        if 'Typical Price'.lower() in label.text.lower() or 'List
Price'.lower() in label.text.lower():

            if 'Typical Price'.lower() in label.text.lower() or 'List
Price'.lower() in label.text.lower():
                typicalPriceText = label.text.strip()
                if ':' in typicalPriceText:
                    typicalPrice = typicalPriceText.split(':')
[1].strip()
                else:
                    typicalPrice = typicalPriceText.split(' ')[-
1].strip()
            break
```

```
    # Extract product details
    details = {}
    detail_bullets = soup.find('table', {'class': 'a-normal a-spacing-
micro'})
    if detail_bullets:
        for tr in detail_bullets.find_all('tr'):
            key = tr.find('td', {'class': 'a-
span3'}).text.strip().replace(':', '').replace('&lrm;','')
            value = tr.find('td', {'class': 'a-
span9'}).text.strip().replace('\u200e', '').replace(':', '')
            details[key] = value

    # Extract technical details
    tech_details = {}
    tech_table = soup.find('table', {'id':
'productDetails_techSpec_section_1'})
    if tech_table:
        for row in tech_table.find_all('tr'):
            key = row.find('th').text.strip().replace('&lrm;','')
            value = row.find('td').text.strip().replace('\u200e',
'').replace(':', '')
            tech_details[key] = value

    # Extract other technical details
    other_tech_details = {}
    other_tech_table = soup.find('table', {'id':
'productDetails_techSpec_section_2'})
    if other_tech_table:
        for row in other_tech_table.find_all('tr'):
            key = row.find('th').text.strip().replace('&lrm;','')
            value = row.find('td').text.strip().replace('\u200e',
'').replace(':', '')
            other_tech_details[key] = value

    # Extract Additional details
    additional_details = {}
    additional_details_table = soup.find('table', {'id':
'productDetails_detailBullets_sections1'})
    if additional_details_table:
        for row in additional_details_table.find_all('tr'):
            key = row.find('th').text.strip().replace('&lrm;','')
            value = row.find('td').text.strip().replace('\u200e',
'').replace(':', '')
            additional_details[key] = value

    #Update - New Amazon UI Structure
    newProdDetails = {}
    newProdDetailsTableList = soup.find_all('table', class_ = 'a-
keyvalue prodDetTable')
    if newProdDetailsTableList:
```

```
        for newProdDetailsTable in newProdDetailsTableList:
            for row in newProdDetailsTable.find_all('tr'):
                key = row.find('th').text.strip().replace('&lrm;','')
                value = row.find('td').text.strip().replace('\u200e',
'').replace(':', '')
                newProdDetails[key] = value

    # Combine all information
    laptop_info = {
        'Title': title,
        'Price': price,
        'Rating': rating,
        'Product Details': details,
        'Technical Details': tech_details,
        'Typical Price': typicalPrice,
        'Additional Details': additional_details,
        'Other Technical Details': other_tech_details,
        'New Product Details': newProdDetails,
        'URL': url
    }

    return laptop_info

#Since we constructed a nested map, we will flatten the keys.
def flatten_dict(d, parent_key='', sep='_'):
    items = []
    for k, v in d.items():
        new_key = f"{parent_key}{sep}{k}" if parent_key else k
        if isinstance(v, dict):
            items.extend(flatten_dict(v, new_key, sep=sep).items())
        else:
            items.append((new_key, v))
    return dict(items)

def process_all_urls():
    new_data = []
    try:
        # Read the file with all Product URLs
        with open('laptop_urls.txt', 'r') as file:
            urls = file.read().splitlines()

        file_exists = os.path.exists(output_file)

        # These are headers(product detail keys) received as per
multiple iterations.
        # But since the Amazon UI structure is dynamic and kept
changing for multiple products/brands/segments, the script is designed
to combine new found headers with existing ones(per scraping batch).
        # So after scraping of each batch the generated csv file can
potentially have different headers.
```

```
        # We won't worry about, we can handle this in the cleaning
phase.
        existing_headers = ['Additional Details_ASIN', 'Additional
Details_Batteries', 'Additional Details_Batteries required',
'Additional Details_Best Sellers Rank', 'Additional Details_Customer
Reviews', 'Additional Details_Date First Available', 'Additional
Details_Form Factor', 'Additional Details_Graphics Card Ram Size',
'Additional Details_Hard Drive Size', 'Additional Details_Included
Components', 'Additional Details_Is Discontinued By Manufacturer',
'Additional Details_Item Weight', 'Additional Details_Item model
number', 'Additional Details_Manufacturer', 'Additional Details_Number
of Ports', 'Additional Details_Processor Speed', 'Additional
Details_Product Dimensions', 'Additional Details_Ram Memory Installed
Size', 'Additional Details_Resolution', 'Additional Details_Scanner
Resolution', 'Additional Details_Specific instructions for use',
'Additional Details_Standing screen display size', 'Additional
Details_Total Usb Ports', 'Additional Details_Warranty Description',
'Other Technical Details_Audio-out Ports (#)', 'Other Technical
Details_Batteries', 'Other Technical Details_Brand', 'Other Technical
Details_Color', 'Other Technical Details_Computer Memory Type', 'Other
Technical Details_Flash Memory Size', 'Other Technical Details_Hard
Drive Interface', 'Other Technical Details_Hard Drive Rotational
Speed', 'Other Technical Details_Hardware Platform', 'Other Technical
Details_Item Dimensions  LxWxH', 'Other Technical Details_Item
Weight', 'Other Technical Details_Item model number', 'Other Technical
Details_Number of Processors', 'Other Technical Details_Operating
System', 'Other Technical Details_Optical Drive Type', 'Other
Technical Details_Package Dimensions', 'Other Technical Details_Power
Source', 'Other Technical Details_Processor Brand', 'Other Technical
Details_Product Dimensions', 'Other Technical Details_Rear Webcam
Resolution', 'Other Technical Details_Series', 'Other Technical
Details_Voltage', 'Price', 'Product Details_Battery Cell Composition',
'Product Details_Brand', 'Product Details_CPU Model', 'Product
Details_CPU Speed', 'Product Details_Cache Size', 'Product
Details_Color', 'Product Details_Connectivity Technology', 'Product
Details_Display Resolution Maximum', 'Product Details_Display
resolution', 'Product Details_Graphics Card Description', 'Product
Details_Graphics Coprocessor', 'Product Details_Graphics Processor
Manufacturer', 'Product Details_Hard Disk Description', 'Product
Details_Hard Disk Size', 'Product Details_Has webcam capability?',
'Product Details_Human Interface Input', 'Product Details_Item
Weight', 'Product Details_Lithium Battery Energy Content', 'Product
Details_Manufacturer', 'Product Details_Memory Slots Available',
'Product Details_Memory Storage Capacity', 'Product Details_Model
Name', 'Product Details_Operating System', 'Product Details_Processor
Count', 'Product Details_RAM Memory Technology', 'Product Details_RAM
Type', 'Product Details_Ram Memory Installed Size', 'Product
Details_Resolution', 'Product Details_Screen Size', 'Product
Details_Special Feature', 'Product Details_Specific Uses For Product',
```

```
'Product Details_Total USB Ports', 'Product Details_Wireless
Communication Technology', 'Rating', 'Technical Details_ASIN',
'Technical Details_Average Battery Life (in hours)', 'Technical
Details_Batteries', 'Technical Details_Card Description', 'Technical
Details_Chipset Brand', 'Technical Details_Country of Origin',
'Technical Details_Date First Available', 'Technical Details_Graphics
Card Ram Size', 'Technical Details_Graphics Coprocessor', 'Technical
Details_Hard Drive', 'Technical Details_Item Weight', 'Technical
Details_Item model number', 'Technical Details_Manufacturer',
'Technical Details_Max Screen Resolution', 'Technical Details_Memory
Speed', 'Technical Details_National Stock Number', 'Technical
Details_Number of USB 2.0 Ports', 'Technical Details_Number of USB 3.0
Ports', 'Technical Details_Processor', 'Technical Details_Product
Dimensions', 'Technical Details_RAM', 'Technical Details_Screen
Resolution', 'Technical Details_Standing screen display size',
'Technical Details_Wireless Type', 'Title', 'Typical Price',
'URL','New Product Details_Keyboard Layout', 'New Product
Details_Control Method', 'New Product Details_Keyboard Description',
'New Product Details_Human-Interface Input', 'New Product
Details_Total Thunderbolt Ports', 'New Product Details_Total Number of
HDMI Ports', 'New Product Details_Number of Ports', 'New Product
Details_Number of Ethernet Ports', 'New Product Details_Total Usb
Ports', 'New Product Details_Ram Memory Maximum Size', 'New Product
Details_RAM Memory Slot Total Count', 'New Product Details_RAM Type',
'New Product Details_RAM Memory Technology', 'New Product Details_RAM
Memory Installed', 'New Product Details_Bluetooth Version', 'New
Product Details_Bluetooth support?', 'New Product Details_Wi-Fi
Generation', 'New Product Details_Wireless Compability', 'New Product
Details_Connectivity Technology', 'New Product Details_Wireless
Technology', 'New Product Details_Graphics Ram Type', 'New Product
Details_Item Dimensions L x W x Thickness', 'New Product
Details_Chipset Type', 'New Product Details_Optical Storage Device',
'New Product Details_Power Device', 'New Product Details_Number of
Drivers', 'New Product Details_Video Output', 'New Product
Details_Virtual Reality Ready', 'New Product Details_Specific Uses For
Product', 'New Product Details_Webcam Capability', 'New Product
Details_Automatic Backup Software Included', 'New Product Details_Form
Factor', 'New Product Details_Hard Disk Interface', 'New Product
Details_Camera Description', 'New Product Details_Color', 'New Product
Details_Hard-Drive Size', 'New Product Details_Operating System', 'New
Product Details_Additional Features', 'New Product Details_Graphics
Description', 'New Product Details_Graphics Coprocessor', 'New Product
Details_Hard Disk Description', 'New Product Details_Video Processor',
'New Product Details_Series Number', 'New Product Details_UPC', 'New
Product Details_Customer Reviews', 'New Product Details_Best Sellers
Rank', 'New Product Details_ASIN', 'New Product Details_Model Number',
'New Product Details_Included Components', 'New Product
Details_Manufacturer', 'New Product Details_Brand Name', 'New Product
Details_Model Name', 'New Product Details_Model Year', 'New Product
```

```
Details_CPU Model Speed Maximum', 'New Product Details_CPU Model
Generation', 'New Product Details_Processor Count', 'New Product
Details_Processor Brand', 'New Product Details_CPU Model Number', 'New
Product Details_Processor Series', 'New Product Details_Processor
Speed', 'New Product Details_Battery Average Life Standby', 'New
Product Details_Battery Average Life', 'New Product Details_Battery
Cell Type', 'New Product Details_Has Color Screen', 'New Product
Details_Screen Finish', 'New Product Details_Supported Monitor Maximum
Quantity', 'New Product Details_Display Type', 'New Product
Details_Display Resolution Maximum', 'New Product Details_Display
Technology', 'New Product Details_Screen Size', 'New Product
Details_Resolution', 'New Product Details_Native Resolution', 'New
Product Details_Audio features', 'New Product Details_Audio
Recording', 'New Product Details_Speaker Description', 'New Product
Details_Microphone Form Factor', 'New Product Details_Audio Output
Type']

        all_headers = set(existing_headers)

        # Process all URLs first to collect all possible headers and
add data to an array.
        for index, url in enumerate(urls[start_index:end_index + 1],
start=start_index):
            print(f"Processing URL # {index} : {url}")
            info = scrape_laptop_info(url)
            if info:
                flat_info = flatten_dict(info)
                all_headers.update(flat_info.keys())
                new_data.append(flat_info)
            else:
                print(f"No data found for URL # {index}")

                with open('final_unprocessed_urls2.txt', 'a') as
unprocessed_file:
                    unprocessed_file.write(f"{url}\n")

            time.sleep(random.uniform(0, 2))


    except FileNotFoundError:
        print("No laptop_urls.txt file found.")
        return
    finally: # Finally block incase of script execution interruption
or failure, we write data fetched so far.
        # If new headers were found or file doesn't exist,
write/update the header row
        if not file_exists or new_data:
            mode = 'w' if not file_exists else 'r+'
            with open(output_file, mode, newline='', encoding='utf-8')
```

```
as csvfile:
                # Convert all_headers to a list and sort it for
consistency
                all_headers = sorted(list(all_headers))
                writer = csv.DictWriter(csvfile,
fieldnames=all_headers)

                if not file_exists:
                    writer.writeheader()

                if file_exists:
                    csvfile.seek(0, 2)

                for row in new_data:
                    complete_row = {header: row.get(header, '') for
header in all_headers}
                    writer.writerow(complete_row)

# Call the function to process all URLs from the .txt file
process_all_urls()
```

Data Cleaning

```
import pandas as pd
import numpy as np

masterDataFilePath = './final_laptop_data.csv'

masterDF = pd.read_csv(masterDataFilePath)

# After the 1st scraping, Amazon changed their UI structure, had to
scrape again for some products with naN values.
# Each file has different headers(keys of product specs)
unprocessedDf1 = pd.read_csv('./final_laptop_data_unprocessed_3.csv')
unprocessedDf2 = pd.read_csv('./final_laptop_data_unprocessed_4.csv')

df = pd.concat([masterDF, unprocessedDf1,unprocessedDf2],
ignore_index=True, sort=False)

# Set the display options
pd.set_option('display.max_columns', None)
pd.set_option('display.max_colwidth', None)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('future.no_silent_downcasting', True)

df.head()
```

```
  Additional Details_ASIN Additional Details_Batteries Additional
Details_Batteries required
Additional Details_Best Sellers Rank
```

Additional Details_Customer Reviews Additional Details_Date First Available Additional Details_Form Factor Additional Details_Graphics Card Ram Size Additional Details_Hard Drive Size Additional Details_Included Components Additional Details_Is Discontinued By Manufacturer Additional Details_Item Weight Additional Details_Item model number Additional Details_Manufacturer Additional Details_Number of Ports Additional Details_Processor Speed Additional Details_Product Dimensions Additional Details_Ram Memory Installed Size Additional Details_Resolution Additional Details_Scanner Resolution Additional Details_Specific instructions for use Additional Details_Standing screen display size Additional Details_Total Usb Ports Additional Details_Warranty Description Other Technical Details_Audio-out Ports (#) Other Technical Details_Batteries Other Technical Details_Brand Other Technical Details_Color Other Technical Details_Computer Memory Type Other Technical Details_Flash Memory Size Other Technical Details_Hard Drive Interface Other Technical Details_Hard Drive Rotational Speed Other Technical Details_Hardware Platform Other Technical Details_Item Dimensions LxWxH Other Technical Details_Item Weight Other Technical Details_Item model number Other Technical Details_Number of Processors Other Technical Details_Operating System Other Technical Details_Optical Drive Type Other Technical Details_Package Dimensions Other Technical Details_Power Source Other Technical Details_Processor Brand Other Technical Details_Product Dimensions Other Technical Details_Rear Webcam Resolution Other Technical Details_Series Other Technical Details_Voltage Price Product Details_Battery Cell Composition Product Details_Brand Product Details_CPU Model Product Details_CPU Speed Product Details_Cache Size Product Details_Color Product Details_Connectivity Technology Product Details_Display Resolution Maximum Product Details_Display resolution Product Details_Graphics Card Description Product Details_Graphics Coprocessor Product Details_Graphics Processor Manufacturer Product Details_Hard Disk Description Product Details_Hard Disk Size Product Details_Has webcam capability? Product Details_Human Interface Input Product Details_Item Weight Product Details_Lithium Battery Energy Content Product Details_Manufacturer Product Details_Memory Slots Available Product Details_Memory Storage Capacity Product Details_Model Name Product Details_Operating System Product Details_Processor Count Product Details_RAM Memory Technology Product Details_RAM Type Product Details_Ram Memory Installed Size Product Details_Resolution Product Details_Screen Size Product Details_Special Feature Product Details_Specific Uses For Product Product Details_Total USB Ports Product Details_Wireless Communication Technology Rating Technical Details_ASIN Technical Details_Average Battery Life (in hours) Technical Details_Batteries Technical Details_Card Description Technical Details_Chipset Brand Technical Details_Country of Origin Technical Details_Date First Available Technical Details_Graphics Card Ram Size Technical Details_Graphics Coprocessor Technical Details_Hard Drive Technical Details_Item Weight

Technical Details_Item model number Technical Details_Manufacturer Technical Details_Max Screen Resolution Technical Details_Memory Speed Technical Details_National Stock Number Technical Details_Number of USB 2.0 Ports Technical Details_Number of USB 3.0 Ports Technical Details_Processor Technical Details_Product Dimensions Technical Details_RAM Technical Details_Screen Resolution Technical Details_Standing screen display size Technical Details_Wireless Type Title Typical Price
URL New Product Details_ASIN New Product Details_Additional Features New Product Details_Audio Output Type New Product Details_Audio Recording New Product Details_Audio features New Product Details_Automatic Backup Software Included New Product Details_Battery Average Life New Product Details_Battery Average Life Standby New Product Details_Battery Cell Type New Product Details_Best Sellers Rank New Product Details_Bluetooth Version New Product Details_Bluetooth support? New Product Details_Brand Name New Product Details_CPU Model Generation New Product Details_CPU Model Number New Product Details_CPU Model Speed Maximum New Product Details_Camera Description New Product Details_Chipset Type New Product Details_Color New Product Details_Connectivity Technology New Product Details_Control Method New Product Details_Customer Reviews New Product Details_Display Resolution Maximum New Product Details_Display Technology New Product Details_Display Type New Product Details_Form Factor New Product Details_Graphics Coprocessor New Product Details_Graphics Description New Product Details_Graphics Ram Type New Product Details_Hard Disk Description New Product Details_Hard Disk Interface New Product Details_Hard-Drive Size New Product Details_Has Color Screen New Product Details_Human-Interface Input New Product Details_Included Components New Product Details_Item Dimensions L x W x Thickness New Product Details_Keyboard Description New Product Details_Keyboard Layout New Product Details_Manufacturer New Product Details_Microphone Form Factor New Product Details_Model Name New Product Details_Model Number New Product Details_Model Year New Product Details_Native Resolution New Product Details_Number of Drivers New Product Details_Number of Ethernet Ports New Product Details_Number of Ports New Product Details_Operating System New Product Details_Optical Storage Device New Product Details_Power Device New Product Details_Processor Brand New Product Details_Processor Count New Product Details_Processor Series New Product Details_Processor Speed New Product Details_RAM Memory Installed New Product Details_RAM Memory Slot Total Count New Product Details_RAM Memory Technology New Product Details_RAM Type New Product Details_Ram Memory Maximum Size New Product Details_Resolution New Product Details_Screen Finish New Product Details_Screen Size New Product Details_Series Number New Product Details_Speaker Description New Product Details_Specific Uses For Product New Product Details_Supported Monitor Maximum Quantity New Product Details_Total Number of HDMI Ports New Product Details_Total Thunderbolt Ports New Product Details_Total Usb Ports New Product Details_UPC New Product

```
Details_Video Output New Product Details_Video Processor New Product
Details_Virtual Reality Ready New Product Details_Webcam Capability
New Product Details_Wi-Fi Generation New Product Details_Wireless
Compability New Product Details_Wireless Technology New Product
Details_Age Range Description  New Product Details_Aspect Ratio  New
Product Details_Available M2 Slot Count New Product Details_Batteries
New Product Details_Battery Capacity New Product Details_Battery Power
New Product Details_Biometric Security Feature New Product
Details_Brand New Product Details_CPU Codename New Product Details_CPU
L3 Cache New Product Details_Cache Memory Installed Size New Product
Details_Card Description New Product Details_Cellular Technology New
Product Details_Chipset Brand New Product Details_Date First Available
New Product Details_Display Refresh Rate in Hertz New Product
Details_Flash Memory Size New Product Details_Front Photo Sensor
Resolution New Product Details_Generation  New Product Details_Global
Trade Identification Number New Product Details_Graphics Card Ram New
Product Details_Hard Disk Rotational Speed New Product Details_Hard
Drive New Product Details_Hard Drive Interface New Product
Details_Hard Drive Rotational Speed New Product Details_Hardware
Connectivity New Product Details_Hardware Interface New Product
Details_Is Electric New Product Details_Item Dimensions  LxWxH New
Product Details_Item Weight New Product Details_Item model number New
Product Details_LAN Port Bandwidth New Product Details_Lithium-Battery
Energy Content New Product Details_Max Screen Resolution New Product
Details_Maximum Display Brightness New Product Details_Memory Clock
Speed New Product Details_Memory Slots Available New Product
Details_Memory Speed New Product Details_Memory Storage Capacity New
Product Details_Notebook Pointing Device Description  New Product
Details_Number Of Cells  New Product Details_Number of Processors  New
Product Details_Number of Rear Facing Cameras  New Product
Details_Number of USB 2.0 Ports  New Product Details_Number of USB 3.0
Ports New Product Details_Optical Drive Type New Product Details_Photo
Sensor Resolution New Product Details_Processor New Product
Details_Processor Description New Product Details_Product Dimensions
New Product Details_RAM New Product Details_Rear Facing Camera Photo
Sensor Resolution New Product Details_Refresh Rate New Product
Details_Screen Bezel Thickness New Product Details_Sensor Type New
Product Details_Series New Product Details_Standing screen display
size New Product Details_Style Number  New Product Details_Total PCIe
Ports New Product Details_Touch Screen Type New Product
Details_Touchpad Feature New Product Details_Video Capture Resolution
New Product Details_Voltage New Product Details_Warranty Type New
Product Details_Wireless Communication Technology
0              B0D8VSDCMK                            NaN
NaN  #81,691 in Computers & Accessories (See Top 100 in Computers &
Accessories)  #16,532 in Traditional Laptop Computers    5.0  5.0 out
of 5 stars    \n    1 rating   \n\n\n 5.0 out of 5 stars
July 5, 2024                               NaN
NaN                                  NaN
```

NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN 1 Lithium Polymer batteries required. (included) ZHAOHUIXIN
Blue NaN
64 GB NaN
NaN NaN
9.53 x 6.89 x 0.83 inches 2.97 pounds
NaN 4.0
Android BD-R
NaN NaN
Alwinner 9.53 x 6.89 x 0.83 inches
NaN PC1068 NaN
119.99 NaN ZHAOHUIXIN
NaN NaN NaN
NaN NaN
1280x800 Pixels NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN 2 GB PC1068
NaN NaN
NaN NaN NaN
NaN 10.1 Inches NaN
NaN NaN
NaN 4.5 out of 5 stars NaN
NaN NaN
Integrated ARM
NaN NaN
NaN NaN 64 GB
Emmc NaN NaN
NaN 1280x800 Pixels
NaN NaN
1.0 1.0 1.8 GHz
a13 NaN DDR4
NaN 10.1 Inches
NaN Mini Android 12 Laptop Computer, Portable Small Netbook with Allwinner A133 CPU Android 12 OS 2GB RAM 64GB EMMC HD IPS Screen 1920x800 0.3MP Camera (Blue)
NaN https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzODc3MDQzMzg2MDI6OjA6Og&url=%2FZHAOHUIXIN-Computer-Portable-

```
Allwinner-1920x800%2Fdp%2FB0D8VSDCMK%2Fref%3Dsr_1_1_sspa%3Fdib
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-1-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1                      NaN
NaN                                   NaN
NaN                                NaN
NaN                                      NaN
NaN                                   NaN
NaN                                    NaN
NaN                            NaN
NaN                                  NaN
NaN                                    NaN
NaN                       NaN
NaN                                NaN
NaN                                            NaN
NaN                              NaN                             NaN
NaN                                      NaN
NaN                                       NaN
NaN                                 NaN
NaN                                       NaN
NaN                                                   NaN
NaN                                 NaN
NaN                                        NaN
NaN                              NaN                             NaN
NaN                                    NaN
NaN                                  NaN
NaN                                        NaN
NaN                                 NaN
NaN                                  NaN
NaN                                      NaN
NaN                                       NaN
NaN                                         NaN
NaN                               NaN                             NaN
NaN                                     NaN
NaN                                                     NaN
NaN                                          NaN
NaN                      NaN                              NaN
NaN                                       NaN
NaN                                  NaN
NaN                                     NaN
NaN                               NaN
NaN                           NaN                                  NaN
NaN                                            NaN
NaN                              NaN                              NaN
```

```
NaN                                  NaN
NaN                               NaN
NaN                                             NaN
NaN                                            NaN
NaN                                                  NaN
NaN                                         NaN
NaN                                   NaN
NaN                                    NaN
NaN                             NaN
NaN                             NaN
NaN                                 NaN
NaN                                    NaN
NaN                                 NaN
NaN                              NaN
NaN                                                   NaN
NaN                                    NaN
NaN                                       NaN
NaN                                 NaN
NaN                           NaN
NaN                                 NaN                     NaN
NaN                              NaN
NaN                             NaN                        NaN
NaN                              NaN
NaN                                   NaN
NaN                                       NaN
NaN                               NaN
NaN
1              B0D87RK5Q8                          NaN
NaN      #5,653 in Computers & Accessories (See Top 100 in Computers &
Accessories)  #670 in Traditional Laptop Computers  4.1  4.1 out of 5
stars    \n    13 ratings   \n\n\n 4.1 out of 5 stars
June 27, 2024                           NaN
NaN                                NaN
NaN                                             NaN
NaN                                  NaN
NaN                                 NaN
NaN                                   NaN
NaN                           NaN
NaN                                          NaN
NaN                                 NaN
NaN                                       NaN  1 Lithium Polymer
batteries required. (included)                          TPV
Silver                                  DDR3 SDRAM
128 MB                                         NaN
NaN                                     NaN
14.09 x 8.97 x 0.86 inches                          5.15 pounds
NaN                                         1.0
Windows 11 Pro                                         NaN
NaN                                  NaN
```

Intel 14.09 x 8.97 x 0.86 inches
NaN AceBook 7.6 Volts
309.99 NaN TPV
Core i5 NaN NaN
Silver NaN
NaN NaN
Integrated NaN
NaN NaN 512
GB NaN
NaN NaN
NaN NaN
NaN NaN AceBook
Windows 11 Pro NaN
NaN NaN 16 GB
NaN 15.6 Inches Webcam
NaN NaN
NaN 4.5 out of 5 stars NaN
5 Hours NaN
Integrated Intel
NaN NaN
NaN Intel UHD Graphics 617 512 GB
SSD NaN NaN
NaN 1920x1080 Pixels
NaN NaN
NaN 2.0 3.6 GHz
core_i5 NaN 16 GB LPDDR3
1920 x 1080 pixels 15.6 Inches
802.11a/b/g/n/ac TPV 15.6" Laptop Computer (Intel Core i5 / 16GB RAM/ 512GB SSD), MS Office 2024, FHD Display with 100% sRGB Color Gamut, Windows 11 Pro Notebook PC with Dual Band Wi-Fi, Webcam (Silver) $369.99 https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzMzE1MDUzNjc5MDI6OjA6Og&url=%2FTPV-Computer-Display-Windows-Notebook%2Fdp%2FB0D87RK5Q8%2Fref%3Dsr_1_2_sspa%3Fdib%3DeyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-2-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN

NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	NaN
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	
NaN	NaN	

NaN NaN
NaN NaN
NaN NaN
NaN NaN NaN
NaN NaN
NaN NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN
2 B0DFWGCCBZ NaN
NaN #46,159 in Computers & Accessories (See Top 100 in Computers & Accessories) #8,150 in Traditional Laptop Computers 5.0 5.0 out of 5 stars \n 5 ratings \n\n\n 5.0 out of 5 stars
August 29, 2024 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN HP NaN
DDR5 RAM NaN
PCIE x 4 NaN
PC 14.14 x 9.2 x 0.78 inches
5 pounds 660 G11
12.0 Windows 11 Pro
No Optical Drive NaN
NaN Intel 14.14 x 9.2 x 0.78 inches NaN
Elitebook NaN 1079.00
NaN HP Intel Core i7
NaN NaN NaN
NaN NaN
NaN Integrated Intel Graphics NaN
NaN 1 TB
NaN NaN NaN
NaN NaN
NaN NaN Elitebook Windows 11 Pro NaN
NaN NaN 32 GB
NaN 16 Inches HD Audio
NaN NaN
NaN 4.0 out of 5 stars NaN

NaN NaN
Integrated Intel
NaN NaN
NaN Intel Graphics 1 TB
SSD NaN NaN
NaN 1920 x 1200 Pixels
NaN NaN
NaN 2.0 1.3 GHz
apple_ci7 NaN DDR5
NaN 16 Inches
NaN HP Elitebook 660 G11 16" FHD+ Business AI Laptop Computer, Intel Ultra 7 155U (Beat i7-1355U), 32GB DDR5 RAM, 1TB PCIe SSD, WiFi 6E, Backlit Keyboard, Fingerprint Reader, Windows 11 Pro, venthear
NaN https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzODcyNTQ1NzUwMDI6OjA6Og&url=%2FHP-Elitebook-660-G11-Fingerprint%2Fdp%2FB0DFWDQS11%2Fref%3Dsr_1_3_sspa%3Fdib%3DeyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-3-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN

```
NaN                                       NaN
NaN                                         NaN
NaN                               NaN                             NaN
NaN                                     NaN
NaN                                                     NaN
NaN                                          NaN
NaN                      NaN                              NaN
NaN                                       NaN
NaN                                  NaN
NaN                                     NaN
NaN                               NaN
NaN                           NaN                                  NaN
NaN                                            NaN
NaN                              NaN                              NaN
NaN                                  NaN
NaN                               NaN
NaN                                                NaN
NaN                                               NaN
NaN                                                     NaN
NaN                                            NaN
NaN                                      NaN
NaN                                       NaN
NaN                             NaN
NaN                             NaN
NaN                                    NaN
NaN                                       NaN
NaN                                    NaN
NaN                              NaN
NaN                                                      NaN
NaN                                       NaN
NaN                                          NaN
NaN                                    NaN
NaN                           NaN
NaN                                    NaN                     NaN
NaN                              NaN
NaN                             NaN                        NaN
NaN                              NaN
NaN                                   NaN
NaN                                          NaN
NaN                               NaN
NaN
3                     NaN                          NaN
NaN
NaN
NaN                                     NaN
NaN                                       NaN
NaN                                     NaN
NaN                            NaN
NaN                             NaN
```

NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN 929.00
NaN Apple Apple M3
NaN NaN Midnight
NaN NaN
NaN Integrated
NaN NaN
NaN 256 GB
NaN NaN NaN
NaN NaN
NaN NaN MacBook Air
Mac OS NaN
NaN NaN 8 GB
NaN 13.6 Inches Fingerprint Reader
NaN NaN
NaN 4.0 out of 5 stars B0CX23V2ZK
NaN 1 Lithium Polymer batteries required. (included)
NaN NaN
China March 4, 2024
NaN NaN
NaN 2.7 pounds MRXV3LL/A
Apple NaN
NaN NaN
NaN NaN
NaN 11.97 x 8.46 x 0.44 inches NaN
NaN NaN
NaN Apple 2024 MacBook Air 13-inch Laptop with M3 chip: Built for Apple Intelligence, 13.6-inch Liquid Retina Display, 8GB Unified Memory, 256GB SSD Storage, Backlit Keyboard, Touch ID; Midnight
$1,099.00
https://www.amazon.com/Apple-2024-MacBook-13-inch-Laptop/dp/B0CX23V2ZK/ref=sr_1_4?dib=eyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY

```
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-
1DtQHxgmY&dib_tag=se&qid=1726997913&s=pc&sr=1-4
NaN                                        NaN
NaN                                    NaN
NaN                                                    NaN
NaN                                                 NaN
NaN                                      NaN
NaN                                       NaN
NaN                                         NaN
NaN                                            NaN
NaN                                  NaN                          NaN
NaN                                    NaN
NaN                                                 NaN
NaN                                  NaN                                  NaN
NaN                                          NaN
NaN                                           NaN
NaN                                    NaN
NaN                                           NaN
NaN                                                    NaN
NaN                                    NaN
NaN                                           NaN
NaN                                  NaN                                  NaN
NaN                                        NaN
NaN                                      NaN
NaN                                           NaN
NaN                                    NaN
NaN                                     NaN
NaN                                         NaN
NaN                                          NaN
NaN                                      NaN                                  NaN
NaN                                            NaN
NaN                                                          NaN
NaN                                                             NaN
NaN                          NaN                                      NaN
NaN                                        NaN
NaN                                     NaN
NaN                                       NaN
NaN                                  NaN                                  NaN
NaN                                                 NaN
NaN                                  NaN                                  NaN
NaN                                      NaN
NaN                                   NaN
NaN                                                       NaN
NaN                                                    NaN
NaN                                                          NaN
NaN                                                 NaN
```

```
NaN                                          NaN
NaN                                           NaN
NaN                                  NaN
NaN                                  NaN
NaN                                         NaN
NaN                                            NaN
NaN                                         NaN
NaN                                   NaN
NaN                                                                  NaN
NaN                                            NaN
NaN                                               NaN
NaN                                         NaN
NaN                                 NaN
NaN                                         NaN                          NaN
NaN                                    NaN
NaN                                   NaN                                NaN
NaN                                    NaN
NaN                                         NaN
NaN                                                 NaN
NaN                                   NaN
NaN
4                          NaN                              NaN
NaN
NaN
NaN                                          NaN
NaN                                            NaN
NaN                                          NaN
NaN                                 NaN
NaN                                  NaN
NaN                                     NaN
NaN                                                 NaN
NaN                                         NaN
NaN                                                    NaN
NaN                                            NaN
NaN                                                       NaN
NaN                                 NaN
NaN                                                  NaN
NaN                                                        NaN
NaN                                                  NaN
NaN                                            NaN
NaN                                          NaN
NaN                                             NaN
NaN                                          NaN
NaN                                                 NaN
NaN                                   NaN  1449.00
NaN                      Apple                       Apple M3
NaN                             NaN                  Midnight
NaN                                                 NaN
NaN                                     Integrated
```

```
NaN                                             NaN
NaN                         512 GB
NaN                                   NaN                         NaN
NaN                          NaN
NaN                                     NaN                  MacBook Air
Mac OS                                NaN
NaN                     NaN                                         16 GB
NaN                15.3 Inches              Fingerprint Reader
NaN                              NaN
NaN  4.0 out of 5 stars            B0CX24BN3L
NaN  1 Lithium Polymer batteries required. (included)
NaN                             NaN
China                          March 4, 2024
NaN                                    NaN
NaN                   3.3 pounds                          MXD43LL/A
Apple                                  NaN
NaN                                     NaN
NaN                                        NaN
NaN           13.4 x 9.35 x 0.45 inches                   NaN
NaN                                            NaN
NaN      Apple 2024 MacBook Air 15-inch Laptop with M3 chip: Built for
Apple Intelligence, 15.3-inch Liquid Retina Display, 16GB Unified
Memory, 512GB SSD Storage, Backlit Keyboard, Touch ID; Midnight
$1,699.00
https://www.amazon.com/Apple-2024-MacBook-15-inch-Laptop/dp/B0CX24BN3L
/ref=sr_1_5?dib=eyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-
1DtQHxgmY&dib_tag=se&qid=1726997913&s=pc&sr=1-5
NaN                                     NaN
NaN                                 NaN
NaN                                                   NaN
NaN                                              NaN
NaN                                   NaN
NaN                                    NaN
NaN                                      NaN
NaN                                         NaN
NaN                              NaN                      NaN
NaN                                NaN
NaN                                            NaN
NaN                              NaN                            NaN
NaN                                      NaN
NaN                                       NaN
NaN                                 NaN
NaN                                       NaN
NaN                                                  NaN
```

```
NaN                                    NaN
NaN                                        NaN
NaN                                  NaN                                    NaN
NaN                                      NaN
NaN                                    NaN
NaN                                        NaN
NaN                                    NaN
NaN                                    NaN
NaN                                         NaN
NaN                                          NaN
NaN                                             NaN
NaN                                  NaN                                    NaN
NaN                                        NaN
NaN                                                          NaN
NaN                                               NaN
NaN                          NaN                                  NaN
NaN                                          NaN
NaN                                      NaN
NaN                                         NaN
NaN                                   NaN
NaN                                NaN                                    NaN
NaN                                          NaN
NaN                                  NaN                                    NaN
NaN                                      NaN
NaN                                    NaN
NaN                                             NaN
NaN                                            NaN
NaN                                                          NaN
NaN                                        NaN
NaN                                                 NaN
NaN                                          NaN
NaN                                    NaN
NaN                                    NaN
NaN                                      NaN
NaN                                        NaN
NaN                                    NaN
NaN                                  NaN                                    NaN
NaN                                                 NaN
NaN                                                    NaN
NaN                                        NaN
NaN                                    NaN
NaN                                  NaN                                    NaN
NaN                                       NaN
NaN                                    NaN                                    NaN
NaN                                      NaN
NaN                                    NaN                                    NaN
NaN                                       NaN
NaN                                             NaN
```

```
NaN                                   NaN
NaN

df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2151 entries, 0 to 2150
Columns: 250 entries, Additional Details_ASIN to New Product
Details_Wireless Communication Technology
dtypes: float64(34), object(216)
memory usage: 4.1+ MB

df.describe()

       Additional Details_Included Components  Additional
Details_Number of Ports  Additional Details_Total Usb Ports
Additional Details_Warranty Description  Other Technical
Details_Audio-out Ports (#)  Other Technical Details_Number of
Processors        Price  Product Details_Processor Count  Product
Details_Total USB Ports  Technical Details_Number of USB 2.0 Ports
Technical Details_Number of USB 3.0 Ports  New Product
Details_Bluetooth Version  New Product Details_Model Year  New Product
Details_Number of Drivers  New Product Details_Number of Ethernet
Ports  New Product Details_Number of Ports  New Product
Details_Processor Count  New Product Details_RAM Memory Slot Total
Count  New Product Details_Series Number  New Product
Details_Supported Monitor Maximum Quantity  New Product Details_Total
Number of HDMI Ports  New Product Details_Total Thunderbolt Ports  New
Product Details_Total Usb Ports  New Product Details_UPC  New Product
Details_Aspect Ratio  New Product Details_Available M2 Slot Count  New
Product Details_Display Refresh Rate in Hertz  New Product
Details_Global Trade Identification Number  New Product Details_Number
Of Cells  New Product Details_Number of Processors  New Product
Details_Number of Rear Facing Cameras  New Product Details_Number of
USB 2.0 Ports  New Product Details_Number of USB 3.0 Ports  New
Product Details_Total PCIe Ports
count                                        4.0
4.000000                               10.000000
4.0                                    10.000000
1939.000000  2150.000000                        13.000000
34.000000                                 533.000000
1342.000000                               28.000000
87.000000                                     2.0
4.00                                39.000000
82.000000                                        23.000000
2.000000                                               22.000000
37.0                                    17.000000
101.000000             8.700000e+01                                2.0
2.0                                               1.0
1.000000e+00                            12.000000
```

```
11.0                                                    1.0
11.000000                                        11.000000
2.0
mean                                          1.0
4.000000                                  2.700000
1.0                                           1.100000
6.326457   690.775247                              7.692308
3.058824                                       1.221388
2.297317                                     5.010714
2022.735632                                          1.0
1.25                                   4.512821
6.134146                                                1.521739
1116.500000                                                        1.454545
1.0                                           1.588235
2.613861             5.381521e+11                                    169.0
1.0                                                   60.0
1.981538e+11                                   4.250000
4.0                                                    1.0
1.090909                                         1.090909
2.0
std                                            0.0
1.414214                                  0.948683
0.0                                           0.316228
4.667055   587.555364                              3.544588
0.422200                                       0.530775
0.960972                                     0.359361
2.137483                                          0.0
0.50                                   1.789910
4.829851                                                0.730477
1225.416052                                                        0.857864
0.0                                           0.507300
0.720698             2.882525e+11                                      0.0
0.0                                                     NaN
NaN                                    4.025487
0.0                                                     NaN
0.301511                                       0.301511
0.0
min                                            1.0
2.000000                                  1.000000
1.0                                           1.000000
1.000000    32.990000                              2.000000
2.000000                                       1.000000
1.000000                                     4.200000
2013.000000                                          1.0
1.00                                   2.000000
1.000000                                                1.000000
250.000000                                                         1.000000
1.0                                           1.000000
1.000000             1.911147e+11                                    169.0
```

```
1.0                                                    60.0
1.981538e+11                                   1.000000
4.0                                                     1.0
1.000000                                          1.000000
2.0
25%                                             1.0
3.500000                                    3.000000
1.0                                             1.000000
2.000000   296.380000                              6.000000
3.000000                                      1.000000
2.000000                                    5.000000
2023.000000                                          1.0
1.00                                    4.000000
2.000000                                               1.000000
683.250000                                                      1.000000
1.0                                             1.000000
2.000000             1.971055e+11                                  169.0
1.0                                                    60.0
1.981538e+11                                    2.750000
4.0                                                     1.0
1.000000                                          1.000000
2.0
50%                                             1.0
4.500000                                    3.000000
1.0                                             1.000000
4.000000   548.715000                              8.000000
3.000000                                      1.000000
2.000000                                    5.100000
2024.000000                                          1.0
1.00                                    5.000000
4.000000                                               1.000000
1116.500000                                                     1.000000
1.0                                             2.000000
3.000000             6.837239e+11                                  169.0
1.0                                                    60.0
1.981538e+11                                    3.000000
4.0                                                     1.0
1.000000                                          1.000000
2.0
75%                                             1.0
5.000000                                    3.000000
1.0                                             1.000000
10.000000   899.987500                              10.000000
3.000000                                      1.000000
3.000000                                    5.200000
2024.000000                                          1.0
1.25                                    6.000000
11.500000                                              2.000000
1549.750000                                                     1.750000
```

```
1.0                                          2.000000
3.000000             7.851831e+11                                169.0
1.0                                                60.0
1.981538e+11                                 4.500000
4.0                                                 1.0
1.000000                                        1.000000
2.0
max                                            1.0
5.000000                                  4.000000
1.0                                          2.000000
64.000000  4939.000000                             16.000000
4.000000                                      4.000000
6.000000                                    5.400000
2024.000000                                         1.0
2.00                                  10.000000
16.000000                                             4.000000
1983.000000                                                     4.000000
1.0                                          2.000000
4.000000             8.898428e+11                                169.0
1.0                                                60.0
1.981538e+11                                 16.000000
4.0                                                 1.0
2.000000                                        2.000000
2.0
```

```
# Create a show summary function since info() was not giving desired
results
def showSummary(df):
    summary = pd.DataFrame({
        'Column': df.columns,
        'Non-Null Count': df.count(),
        'Null Count': df.isnull().sum(),
        'Dtype': df.dtypes,
        'Unique Values': df.nunique(),
    })

    summary['Null Percentage'] = (summary['Null Count'] / len(df)) *
100

    summary = summary[['Column', 'Non-Null Count', 'Null Count', 'Null
Percentage', 'Dtype', 'Unique Values']]

    summary = summary.sort_values('Non-Null Count', ascending=False)

    print(summary.to_string(index=False))
showSummary(df)
```

```
                                                  Column  Non-
Null Count  Null Count  Null Percentage   Dtype  Unique Values
                                                   Title
```

```
2151           0         0.000000  object           1416                                  Rating
2151           0         0.000000  object             21                                     URL
2151           0         0.000000  object           1987                                   Price
2150           1         0.046490 float64            906                  Product Details_Brand
2140          11         0.511390  object             74                 Product Details_Screen Size
2137          14         0.650860  object             49                  Product Details_Model Name
2080          71         3.300790  object            817        Product Details_Ram Memory Installed Size
2012         139         6.462111  object             18             Product Details_Operating System
2007         144         6.694561  object             47        Product Details_Graphics Card Description
2001         150         6.973501  object             34                     Additional Details_ASIN
1995         156         7.252441  object           1393          Additional Details_Date First Available
1991         160         7.438401  object            708                    Product Details_CPU Model
1986         165         7.670851  object            118           Other Technical Details_Brand
1973         178         8.275221  object             73     Technical Details_Standing screen display size
1972         179         8.321711  object             48                 Technical Details_Hard Drive
1959         192         8.926081  object             63                  Technical Details_Processor
1955         196         9.112041  object            508           Additional Details_Best Sellers Rank
1953         198         9.205021  object           1414          Other Technical Details_Operating System
1949         202         9.390981  object             46        Other Technical Details_Product Dimensions
1944         207         9.623431  object            800     Other Technical Details_Item Dimensions  LxWxH
1944         207         9.623431  object            800           Other Technical Details_Processor Brand
1943         208         9.669921  object             16              Technical Details_Card Description
1941         210         9.762901  object             32                Other Technical Details_Item Weight
1941         210         9.762901  object            345      Other Technical Details_Number of Processors
1939         212         9.855881 float64             14
```

```
                                Other Technical Details_Series
1936         215         9.995351  object            811
                               Product Details_Hard Disk Size
1935         216        10.041841  object             24
                                        Technical Details_RAM
1930         221        10.274291  object            148
                      Technical Details_Max Screen Resolution
1870         281        13.063691  object            103
                              Technical Details_Chipset Brand
1869         282        13.110181  object             21
                         Additional Details_Customer Reviews
1726         425        19.758252  object            856
                         Technical Details_Graphics Coprocessor
1708         443        20.595072  object            210
                   Other Technical Details_Computer Memory Type
1695         456        21.199442  object             15
                      Other Technical Details_Item model number
1659         492        22.873082  object            887
                              Product Details_Special Feature
1618         533        24.779172  object            316
                         Technical Details_Screen Resolution
1612         539        25.058113  object             65
                                        Product Details_Color
1604         547        25.430033  object            166
                                 Other Technical Details_Color
1558         593        27.568573  object            166
                   Other Technical Details_Hard Drive Interface
1510         641        29.800093  object             29
                      Technical Details_Number of USB 3.0 Ports
1342         809        37.610414 float64              6
                              Technical Details_Wireless Type
1322         829        38.540214  object             64
                      Other Technical Details_Hardware Platform
1272         879        40.864714  object             10
                           Other Technical Details_Batteries
1104        1047        48.675035  object             22
                    Other Technical Details_Optical Drive Type
997         1154        53.649465  object             41
                    Other Technical Details_Flash Memory Size
945         1206        56.066946  object             38
                               Technical Details_Memory Speed
737         1414        65.736867  object             63
                          Other Technical Details_Power Source
718         1433        66.620177  object              4
                              Other Technical Details_Voltage
650         1501        69.781497  object             70
                     Technical Details_Graphics Card Ram Size
588         1563        72.663877  object             37
                      Technical Details_Number of USB 2.0 Ports
```

```
533        1618        75.220828 float64              4
                      Product Details_Graphics Coprocessor
524        1627        75.639238  object             99
            Technical Details_Average Battery Life (in hours)
483        1668        77.545328  object             69
                                              Typical Price
370        1781        82.798698  object            203
                                 Product Details_CPU Speed
308        1843        85.681079  object             58
          Other Technical Details_Hard Drive Rotational Speed
162        1989        92.468619  object             28
                                  New Product Details_ASIN
146        2005        93.212459  object             82
                        New Product Details_Best Sellers Rank
140        2011        93.491399  object             80
                        Product Details_Hard Disk Description
138        2013        93.584379  object              4
                               New Product Details_Manufacturer
135        2016        93.723849  object             25
                                 New Product Details_Brand Name
133        2018        93.816829  object             19
                         New Product Details_Customer Reviews
132        2019        93.863319  object             61
                                New Product Details_Screen Size
132        2019        93.863319  object             22
                         New Product Details_Operating System
125        2026        94.188749  object             11
                             New Product Details_Video Processor
125        2026        94.188749  object              8
                             New Product Details_Processor Brand
124        2027        94.235239  object              6
                 New Product Details_Display Resolution Maximum
119        2032        94.467689  object             28
                                    New Product Details_Color
119        2032        94.467689  object             28
                       New Product Details_RAM Memory Installed
119        2032        94.467689  object             10
                             New Product Details_Processor Speed
117        2034        94.560669  object             31
                       New Product Details_Graphics Description
117        2034        94.560669  object              4
                                 New Product Details_Model Name
116        2035        94.607159  object             53
                               New Product Details_Display Type
109        2042        94.932589  object              7
          New Product Details_Item Dimensions L x W x Thickness
104        2047        95.165040  object             49
                       New Product Details_Additional Features
103        2048        95.211530  object             33
```

```
                          New Product Details_Total Usb Ports
101        2050        95.304510 float64              4
                       New Product Details_CPU Model Number
101        2050        95.304510  object             43
                       New Product Details_Included Components
99         2052        95.397490  object             45
                    New Product Details_Connectivity Technology
97         2054        95.490470  object             22
                       New Product Details_Battery Cell Type
94         2057        95.629940  object              2
                       Product Details_Memory Storage Capacity
94         2057        95.629940  object              5
                  New Product Details_Specific Uses For Product
93         2058        95.676430  object             23
                     Product Details_Display Resolution Maximum
90         2061        95.815900  object              4
                       New Product Details_Wireless Technology
89         2062        95.862390  object              7
                      New Product Details_Human-Interface Input
88         2063        95.908880  object             20
                       New Product Details_Audio Output Type
87         2064        95.955370  object              9
                                New Product Details_Model Year
87         2064        95.955370 float64              9
                                      New Product Details_UPC
87         2064        95.955370 float64             70
                       New Product Details_Native Resolution
86         2065        96.001860  object             21
                      New Product Details_RAM Memory Technology
84         2067        96.094840  object             14
                              New Product Details_Model Number
84         2067        96.094840  object             63
                           New Product Details_Hard-Drive Size
83         2068        96.141330  object             11
                           New Product Details_Processor Count
82         2069        96.187820 float64             10
                                New Product Details_Resolution
80         2071        96.280800  object             15
                      New Product Details_Hard Disk Description
80         2071        96.280800  object              6
                                 New Product Details_RAM Type
79         2072        96.327290  object              9
                      New Product Details_Graphics Coprocessor
78         2073        96.373780  object             31
                      New Product Details_Keyboard Description
76         2075        96.466760  object             14
                        New Product Details_Camera Description
76         2075        96.466760  object             12
                           New Product Details_Processor Series
```

```
73        2078        96.606230  object             20
                  New Product Details_Ram Memory Maximum Size
68        2083        96.838680  object             11
                         New Product Details_Control Method
67        2084        96.885170  object              2
                     New Product Details_Wireless Compability
67        2084        96.885170  object             16
                      New Product Details_Hard Disk Interface
65        2086        96.978150  object              9
                            New Product Details_Form Factor
60        2091        97.210600  object             13
            New Product Details_Lithium-Battery Energy Content
56        2095        97.396560  object             17
       New Product Details_Automatic Backup Software Included
53        2098        97.536030  object             18
                            New Product Details_Item Weight
53        2098        97.536030  object             29
                               Product Details_Item Weight
52        2099        97.582520  object             31
                         New Product Details_Graphics Ram Type
52        2099        97.582520  object              7
                         New Product Details_Webcam Capability
51        2100        97.629010  object              2
                           New Product Details_Keyboard Layout
51        2100        97.629010  object              1
                         New Product Details_Display Technology
48        2103        97.768480  object              9
                                Product Details_Resolution
48        2103        97.768480  object              8
                  Other Technical Details_Rear Webcam Resolution
48        2103        97.768480  object             10
                            New Product Details_Video Output
46        2105        97.861460  object              9
                     New Product Details_Memory Slots Available
45        2106        97.907950  object              3
                       New Product Details_Bluetooth support?
45        2106        97.907950  object              2
                            New Product Details_Power Device
44        2107        97.954440  object             12
                           New Product Details_Screen Finish
39        2112        98.186890  object             14
                           New Product Details_Number of Ports
39        2112        98.186890 float64              6
                          New Product Details_Has Color Screen
38        2113        98.233380  object              1
                     New Product Details_CPU Model Speed Maximum
37        2114        98.279870  object             14
                  New Product Details_Total Number of HDMI Ports
37        2114        98.279870 float64              1
```

```
                      New Product Details_Microphone Form Factor
35        2116        98.372850  object             11
                    New Product Details_Memory Storage Capacity
34        2117        98.419340  object              3
                     New Product Details_Processor Description
34        2117        98.419340  object              2
                     New Product Details_Age Range Description
34        2117        98.419340  object              1
                       New Product Details_Hardware Interface
34        2117        98.419340  object              3
         New Product Details_Wireless Communication Technology
34        2117        98.419340  object              1
                               Product Details_Total USB Ports
34        2117        98.419340 float64              3
                      New Product Details_CPU Model Generation
34        2117        98.419340  object             18
                            New Product Details_Audio features
31        2120        98.558810  object             24
                           New Product Details_Audio Recording
31        2120        98.558810  object              2
                           New Product Details_Wi-Fi Generation
29        2122        98.651790  object             13
                          New Product Details_Bluetooth Version
28        2123        98.698280 float64              5
                     New Product Details_Hardware Connectivity
28        2123        98.698280  object             21
                    Other Technical Details_Package Dimensions
27        2124        98.744770  object             19
                    New Product Details_Optical Storage Device
26        2125        98.791260  object              5
                       Product Details_Connectivity Technology
25        2126        98.837750  object              8
                      New Product Details_Battery Average Life
25        2126        98.837750  object             15
               New Product Details_RAM Memory Slot Total Count
23        2128        98.930730 float64              3
        New Product Details_Supported Monitor Maximum Quantity
22        2129        98.977220 float64              4
                              New Product Details_Chipset Type
21        2130        99.023710  object             16
                       New Product Details_Speaker Description
21        2130        99.023710  object             14
                                Additional Details_Item Weight
20        2131        99.070200  object             13
                              Additional Details_Manufacturer
20        2131        99.070200  object              8
                             New Product Details_Warranty Type
20        2131        99.070200  object              2
                                   New Product Details_Voltage
```

```
19        2132        99.116690  object             13
                  New Product Details_Total Thunderbolt Ports
17        2134        99.209670 float64              2
               Product Details_Graphics Processor Manufacturer
17        2134        99.209670  object              2
                     Additional Details_Product Dimensions
17        2134        99.209670  object             11
                    New Product Details_Memory Clock Speed
16        2135        99.256160  object              5
                             New Product Details_Is Electric
15        2136        99.302650  object              1
                         Additional Details_Item model number
14        2137        99.349140  object              8
                New Product Details_Biometric Security Feature
14        2137        99.349140  object              3
                         Technical Details_Product Dimensions
13        2138        99.395630  object              7
                               New Product Details_Memory Speed
13        2138        99.395630  object              8
                              Product Details_Processor Count
13        2138        99.395630 float64              6
                           Technical Details_Item model number
13        2138        99.395630  object             13
                             Technical Details_Item Weight
13        2138        99.395630  object             11
               Additional Details_Standing screen display size
13        2138        99.395630  object              6
                     New Product Details_Graphics Card Ram
12        2139        99.442120  object              7
                              Technical Details_Manufacturer
12        2139        99.442120  object              1
                               New Product Details_CPU Codename
12        2139        99.442120  object             10
                    Technical Details_Date First Available
12        2139        99.442120  object              8
             New Product Details_Front Photo Sensor Resolution
12        2139        99.442120  object              4
                          New Product Details_Number Of Cells
12        2139        99.442120 float64              6
                               New Product Details_Refresh Rate
12        2139        99.442120  object              5
                                          Technical Details_ASIN
12        2139        99.442120  object             12
                              New Product Details_Hard Drive
11        2140        99.488610  object              2
                    New Product Details_Product Dimensions
11        2140        99.488610  object              2
                  New Product Details_Number of USB 3.0 Ports
11        2140        99.488610 float64              2
```

```
                      New Product Details_Max Screen Resolution
11        2140        99.488610  object              2
                                 New Product Details_Processor
11        2140        99.488610  object              2
                                     New Product Details_Brand
11        2140        99.488610  object              2
                    New Product Details_Item Dimensions  LxWxH
11        2140        99.488610  object              2
                                       New Product Details_RAM
11        2140        99.488610  object              2
                          New Product Details_Card Description
11        2140        99.488610  object              1
                               New Product Details_Chipset Brand
11        2140        99.488610  object              2
              New Product Details_Standing screen display size
11        2140        99.488610  object              2
                      New Product Details_Date First Available
11        2140        99.488610  object              2
                                    New Product Details_Series
11        2140        99.488610  object              2
                         New Product Details_Flash Memory Size
11        2140        99.488610  object              2
                      New Product Details_Number of Processors
11        2140        99.488610 float64              1
                   New Product Details_Number of USB 2.0 Ports
11        2140        99.488610 float64              2
                                  Additional Details_Batteries
11        2140        99.488610  object              3
                                    Technical Details_Batteries
11        2140        99.488610  object              3
                    Additional Details_Ram Memory Installed Size
11        2140        99.488610  object              5
                        New Product Details_Optical Drive Type
10        2141        99.535100  object              1
                              Additional Details_Processor Speed
10        2141        99.535100  object              7
                              Additional Details_Hard Drive Size
10        2141        99.535100  object              5
      New Product Details_Notebook Pointing Device Description
10        2141        99.535100  object              5
                              Additional Details_Total Usb Ports
10        2141        99.535100 float64              3
                   Other Technical Details_Audio-out Ports (#)
10        2141        99.535100 float64              2
                                 New Product Details_Batteries
10        2141        99.535100  object              1
                      Product Details_Specific Uses For Product
9         2142        99.581590  object              3
                                  Additional Details_Resolution
```

```
9        2142        99.581590  object              5
                         New Product Details_Touchpad Feature
9        2142        99.581590  object              5
                  New Product Details_Hard Disk Rotational Speed
8        2143        99.628080  object              3
                     Additional Details_Graphics Card Ram Size
8        2143        99.628080  object              3
                         Product Details_Human Interface Input
8        2143        99.628080  object              4
                      Product Details_Battery Cell Composition
8        2143        99.628080  object              1
                    New Product Details_Screen Bezel Thickness
7        2144        99.674570  object              6
                       New Product Details_Cellular Technology
7        2144        99.674570  object              3
                            Technical Details_Country of Origin
7        2144        99.674570  object              1
               New Product Details_Battery Average Life Standby
7        2144        99.674570  object              3
                         Product Details_RAM Memory Technology
7        2144        99.674570  object              3
            Additional Details_Is Discontinued By Manufacturer
7        2144        99.674570  object              1
                  New Product Details_Maximum Display Brightness
6        2145        99.721060  object              4
                        Product Details_Has webcam capability?
6        2145        99.721060  object              1
                         Additional Details_Scanner Resolution
6        2145        99.721060  object              2
                         New Product Details_Touch Screen Type
6        2145        99.721060  object              4
                     New Product Details_Virtual Reality Ready
5        2146        99.767550  object              1
                                   Product Details_Manufacturer
5        2146        99.767550  object              3
                  Product Details_Lithium Battery Energy Content
4        2147        99.814040  object              2
                            Additional Details_Number of Ports
4        2147        99.814040 float64              3
                        Product Details_Memory Slots Available
4        2147        99.814040  object              3
New Product Details_Rear Facing Camera Photo Sensor Resolution
4        2147        99.814040  object              2
               Product Details_Wireless Communication Technology
4        2147        99.814040  object              1
                    New Product Details_Number of Ethernet Ports
4        2147        99.814040 float64              2
                                 Additional Details_Form Factor
4        2147        99.814040  object              1
```

```
                          Additional Details_Batteries required
4        2147        99.814040  object              1
             Additional Details_Specific instructions for use
4        2147        99.814040  object              1
                       Additional Details_Warranty Description
4        2147        99.814040 float64              1
                        Additional Details_Included Components
4        2147        99.814040 float64              1
                            Product Details_Display resolution
4        2147        99.814040  object              1
              New Product Details_Cache Memory Installed Size
3        2148        99.860530  object              2
                              New Product Details_CPU L3 Cache
3        2148        99.860530  object              3
                                  Product Details_Cache Size
3        2148        99.860530  object              1
                                    Product Details_RAM Type
3        2148        99.860530  object              2
                              New Product Details_Style Number
2        2149        99.907020  object              2
                       Technical Details_National Stock Number
2        2149        99.907020  object              2
                           New Product Details_Total PCIe Ports
2        2149        99.907020 float64              1
                             New Product Details_Sensor Type
2        2149        99.907020  object              1
                              New Product Details_Aspect Ratio
2        2149        99.907020 float64              1
                    New Product Details_Photo Sensor Resolution
2        2149          99.907020  object              2
                          New Product Details_Number of Drivers
2        2149        99.907020 float64              1
                    New Product Details_Available M2 Slot Count
2        2149          99.907020 float64              1
                                 New Product Details_Series Number
2        2149        99.907020 float64              2
                 New Product Details_Number of Rear Facing Cameras
1        2150        99.953510 float64              1
                         New Product Details_LAN Port Bandwidth
1        2150        99.953510  object              1
                          New Product Details_Item model number
1        2150        99.953510  object              1
               New Product Details_Hard Drive Rotational Speed
1        2150        99.953510  object              1
                        New Product Details_Hard Drive Interface
1        2150        99.953510  object              1
       New Product Details_Global Trade Identification Number
1        2150        99.953510 float64              1
                                  New Product Details_Generation
1        2150        99.953510  object              1
```

```
                    New Product Details_Video Capture Resolution
1        2150        99.953510  object                1
             New Product Details_Display Refresh Rate in Hertz
1        2150        99.953510 float64                1
                             New Product Details_Battery Power
1        2150        99.953510  object                1
                          New Product Details_Battery Capacity
1        2150        99.953510  object                1
```

Most products on amazon have details in an unstructured manner and there is no fixed structure to scrape these details. Thus I have details across multiple columns.

I will combine these scattered product details into 1 column

```
columns = list(df.columns)

# For Brand Name
brandColumnsFiltered = [brand for brand in columns if 'Brand'.lower()
in brand.lower() ]
brandColumnsFiltered
```

```
['Other Technical Details_Brand',
 'Other Technical Details_Processor Brand',
 'Product Details_Brand',
 'Technical Details_Chipset Brand',
 'New Product Details_Brand Name',
 'New Product Details_Processor Brand',
 'New Product Details_Brand',
 'New Product Details_Chipset Brand']
```

```
brandColumns = ['Other Technical Details_Brand','Product
Details_Brand','New Product Details_Brand Name','New Product
Details_Brand']
# Creating the new column with any non-naN value found from the chosen
columns associated to Brand.
df = df.copy()
df['Brand'] = df[brandColumns].bfill(axis=1).iloc[:,0]

# For Processor Brand
processorColumnsFiltered = [processor for processor in columns if
'Processor'.lower() in processor.lower() or 'CPU'.lower() in
processor.lower()]
processorColumnsFiltered
```

```
['Additional Details_Processor Speed',
 'Other Technical Details_Number of Processors',
 'Other Technical Details_Processor Brand',
 'Product Details_CPU Model',
 'Product Details_CPU Speed',
```

```
 'Product Details_Graphics Coprocessor',
 'Product Details_Graphics Processor Manufacturer',
 'Product Details_Processor Count',
 'Technical Details_Graphics Coprocessor',
 'Technical Details_Processor',
 'New Product Details_CPU Model Generation',
 'New Product Details_CPU Model Number',
 'New Product Details_CPU Model Speed Maximum',
 'New Product Details_Graphics Coprocessor',
 'New Product Details_Processor Brand',
 'New Product Details_Processor Count',
 'New Product Details_Processor Series',
 'New Product Details_Processor Speed',
 'New Product Details_Video Processor',
 'New Product Details_CPU Codename',
 'New Product Details_CPU L3 Cache',
 'New Product Details_Number of Processors',
 'New Product Details_Processor',
 'New Product Details_Processor Description']
```

```
processorBrandColumns = ['Other Technical Details_Processor
Brand','New Product Details_Processor Brand']
df = df.copy()
df['Processor_Brand'] =
df[processorBrandColumns].bfill(axis=1).iloc[:,0]

# For Processor Model
processorModelColumns = ['New Product Details_CPU Model
Number','Product Details_CPU Model','Technical Details_Processor','New
Product Details_CPU Codename','New Product Details_Processor']
df = df.copy()
df['Processor_Model'] =
df[processorModelColumns].bfill(axis=1).iloc[:,0]
df.head(2)
```

```
  Additional Details_ASIN Additional Details_Batteries Additional
Details_Batteries required
Additional Details_Best Sellers Rank
Additional Details_Customer Reviews Additional Details_Date First
Available Additional Details_Form Factor Additional Details_Graphics
Card Ram Size Additional Details_Hard Drive Size  Additional
Details_Included Components Additional Details_Is Discontinued By
Manufacturer Additional Details_Item Weight Additional Details_Item
model number Additional Details_Manufacturer  Additional
Details_Number of Ports Additional Details_Processor Speed Additional
Details_Product Dimensions Additional Details_Ram Memory Installed
Size Additional Details_Resolution Additional Details_Scanner
Resolution Additional Details_Specific instructions for use Additional
Details_Standing screen display size  Additional Details_Total Usb
Ports  Additional Details_Warranty Description  Other Technical
```

Details_Audio-out Ports (#) Other Technical Details_Batteries Other Technical Details_Brand Other Technical Details_Color Other Technical Details_Computer Memory Type Other Technical Details_Flash Memory Size Other Technical Details_Hard Drive Interface Other Technical Details_Hard Drive Rotational Speed Other Technical Details_Hardware Platform Other Technical Details_Item Dimensions LxWxH Other Technical Details_Item Weight Other Technical Details_Item model number Other Technical Details_Number of Processors Other Technical Details_Operating System Other Technical Details_Optical Drive Type Other Technical Details_Package Dimensions Other Technical Details_Power Source Other Technical Details_Processor Brand Other Technical Details_Product Dimensions Other Technical Details_Rear Webcam Resolution Other Technical Details_Series Other Technical Details_Voltage Price Product Details_Battery Cell Composition Product Details_Brand Product Details_CPU Model Product Details_CPU Speed Product Details_Cache Size Product Details_Color Product Details_Connectivity Technology Product Details_Display Resolution Maximum Product Details_Display resolution Product Details_Graphics Card Description Product Details_Graphics Coprocessor Product Details_Graphics Processor Manufacturer Product Details_Hard Disk Description Product Details_Hard Disk Size Product Details_Has webcam capability? Product Details_Human Interface Input Product Details_Item Weight Product Details_Lithium Battery Energy Content Product Details_Manufacturer Product Details_Memory Slots Available Product Details_Memory Storage Capacity Product Details_Model Name Product Details_Operating System Product Details_Processor Count Product Details_RAM Memory Technology Product Details_RAM Type Product Details_Ram Memory Installed Size Product Details_Resolution Product Details_Screen Size Product Details_Special Feature Product Details_Specific Uses For Product Product Details_Total USB Ports Product Details_Wireless Communication Technology Rating Technical Details_ASIN Technical Details_Average Battery Life (in hours) Technical Details_Batteries Technical Details_Card Description Technical Details_Chipset Brand Technical Details_Country of Origin Technical Details_Date First Available Technical Details_Graphics Card Ram Size Technical Details_Graphics Coprocessor Technical Details_Hard Drive Technical Details_Item Weight Technical Details_Item model number Technical Details_Manufacturer Technical Details_Max Screen Resolution Technical Details_Memory Speed Technical Details_National Stock Number Technical Details_Number of USB 2.0 Ports Technical Details_Number of USB 3.0 Ports Technical Details_Processor Technical Details_Product Dimensions Technical Details_RAM Technical Details_Screen Resolution Technical Details_Standing screen display size Technical Details_Wireless Type
Title Typical Price
URL New Product Details_ASIN New Product Details_Additional Features New Product Details_Audio Output Type New Product Details_Audio Recording New Product Details_Audio features New Product Details_Automatic Backup Software Included New Product Details_Battery

Average Life New Product Details_Battery Average Life Standby New Product Details_Battery Cell Type New Product Details_Best Sellers Rank New Product Details_Bluetooth Version New Product Details_Bluetooth support? New Product Details_Brand Name New Product Details_CPU Model Generation New Product Details_CPU Model Number New Product Details_CPU Model Speed Maximum New Product Details_Camera Description New Product Details_Chipset Type New Product Details_Color New Product Details_Connectivity Technology New Product Details_Control Method New Product Details_Customer Reviews New Product Details_Display Resolution Maximum New Product Details_Display Technology New Product Details_Display Type New Product Details_Form Factor New Product Details_Graphics Coprocessor New Product Details_Graphics Description New Product Details_Graphics Ram Type New Product Details_Hard Disk Description New Product Details_Hard Disk Interface New Product Details_Hard-Drive Size New Product Details_Has Color Screen New Product Details_Human-Interface Input New Product Details_Included Components New Product Details_Item Dimensions L x W x Thickness New Product Details_Keyboard Description New Product Details_Keyboard Layout New Product Details_Manufacturer New Product Details_Microphone Form Factor New Product Details_Model Name New Product Details_Model Number New Product Details_Model Year New Product Details_Native Resolution New Product Details_Number of Drivers New Product Details_Number of Ethernet Ports New Product Details_Number of Ports New Product Details_Operating System New Product Details_Optical Storage Device New Product Details_Power Device New Product Details_Processor Brand New Product Details_Processor Count New Product Details_Processor Series New Product Details_Processor Speed New Product Details_RAM Memory Installed New Product Details_RAM Memory Slot Total Count New Product Details_RAM Memory Technology New Product Details_RAM Type New Product Details_Ram Memory Maximum Size New Product Details_Resolution New Product Details_Screen Finish New Product Details_Screen Size New Product Details_Series Number New Product Details_Speaker Description New Product Details_Specific Uses For Product New Product Details_Supported Monitor Maximum Quantity New Product Details_Total Number of HDMI Ports New Product Details_Total Thunderbolt Ports New Product Details_Total Usb Ports New Product Details_UPC New Product Details_Video Output New Product Details_Video Processor New Product Details_Virtual Reality Ready New Product Details_Webcam Capability New Product Details_Wi-Fi Generation New Product Details_Wireless Compability New Product Details_Wireless Technology New Product Details_Age Range Description New Product Details_Aspect Ratio New Product Details_Available M2 Slot Count New Product Details_Batteries New Product Details_Battery Capacity New Product Details_Battery Power New Product Details_Biometric Security Feature New Product Details_Brand New Product Details_CPU Codename New Product Details_CPU L3 Cache New Product Details_Cache Memory Installed Size New Product Details_Card Description New Product Details_Cellular Technology New Product Details_Chipset Brand New Product Details_Date First Available

```
New Product Details_Display Refresh Rate in Hertz New Product
Details_Flash Memory Size New Product Details_Front Photo Sensor
Resolution New Product Details_Generation  New Product Details_Global
Trade Identification Number New Product Details_Graphics Card Ram New
Product Details_Hard Disk Rotational Speed New Product Details_Hard
Drive New Product Details_Hard Drive Interface New Product
Details_Hard Drive Rotational Speed New Product Details_Hardware
Connectivity New Product Details_Hardware Interface New Product
Details_Is Electric New Product Details_Item Dimensions  LxWxH New
Product Details_Item Weight New Product Details_Item model number New
Product Details_LAN Port Bandwidth New Product Details_Lithium-Battery
Energy Content New Product Details_Max Screen Resolution New Product
Details_Maximum Display Brightness New Product Details_Memory Clock
Speed New Product Details_Memory Slots Available New Product
Details_Memory Speed New Product Details_Memory Storage Capacity New
Product Details_Notebook Pointing Device Description  New Product
Details_Number Of Cells  New Product Details_Number of Processors  New
Product Details_Number of Rear Facing Cameras  New Product
Details_Number of USB 2.0 Ports  New Product Details_Number of USB 3.0
Ports New Product Details_Optical Drive Type New Product Details_Photo
Sensor Resolution New Product Details_Processor New Product
Details_Processor Description New Product Details_Product Dimensions
New Product Details_RAM New Product Details_Rear Facing Camera Photo
Sensor Resolution New Product Details_Refresh Rate New Product
Details_Screen Bezel Thickness New Product Details_Sensor Type New
Product Details_Series New Product Details_Standing screen display
size New Product Details_Style Number  New Product Details_Total PCIe
Ports New Product Details_Touch Screen Type New Product
Details_Touchpad Feature New Product Details_Video Capture Resolution
New Product Details_Voltage New Product Details_Warranty Type New
Product Details_Wireless Communication Technology       Brand
Processor_Brand Processor_Model
0              B0D8VSDCMK                              NaN
NaN  #81,691 in Computers & Accessories (See Top 100 in Computers &
Accessories)  #16,532 in Traditional Laptop Computers    5.0  5.0 out
of 5 stars    \n    1 rating   \n\n\n 5.0 out of 5 stars
July 5, 2024                                 NaN
NaN                                  NaN
NaN                                                NaN
NaN                                    NaN
NaN                                   NaN
NaN                                     NaN
NaN                             NaN
NaN                                                NaN
NaN                                   NaN
NaN                                             NaN  1 Lithium Polymer
batteries required. (included)                      ZHAOHUIXIN
Blue                                             NaN
64 GB                                             NaN
```

NaN NaN
9.53 x 6.89 x 0.83 inches 2.97 pounds
NaN 4.0
Android BD-R
NaN NaN
Alwinner 9.53 x 6.89 x 0.83 inches
NaN PC1068 NaN
119.99 NaN ZHAOHUIXIN
NaN NaN NaN
NaN NaN
1280x800 Pixels NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN 2 GB PC1068
NaN NaN
NaN NaN NaN
NaN 10.1 Inches NaN
NaN NaN
NaN 4.5 out of 5 stars NaN
NaN NaN Integrated
ARM NaN
NaN NaN
NaN 64 GB Emmc NaN
NaN NaN 1280x800
Pixels NaN
NaN 1.0
1.0 1.8 GHz a13 NaN
DDR4 NaN
10.1 Inches NaN
Mini Android 12 Laptop Computer, Portable Small Netbook with Allwinner A133 CPU Android 12 OS 2GB RAM 64GB EMMC HD IPS Screen 1920x800 0.3MP Camera (Blue) NaN https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzODc3MDQzMzg2MDI6OjA6Og&url=%2FZHAOHUIXIN-Computer-Portable-Allwinner-1920x800%2Fdp%2FB0D8VSDCMK%2Fref%3Dsr_1_1_sspa%3Fdib%3DeyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-1-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1 NaN
NaN NaN
NaN NaN
NaN NaN

```
NaN                                   NaN
NaN                                    NaN
NaN                            NaN
NaN                                  NaN
NaN                                    NaN
NaN                       NaN
NaN                                NaN
NaN                                            NaN
NaN                              NaN                             NaN
NaN                                      NaN
NaN                                       NaN
NaN                                 NaN
NaN                                       NaN
NaN                                                   NaN
NaN                                 NaN
NaN                                        NaN
NaN                              NaN                             NaN
NaN                                    NaN
NaN                                  NaN
NaN                                        NaN
NaN                                 NaN
NaN                                  NaN
NaN                                      NaN
NaN                                       NaN
NaN                                         NaN
NaN                               NaN                             NaN
NaN                                     NaN
NaN                                                     NaN
NaN                                          NaN
NaN                      NaN                              NaN
NaN                                       NaN
NaN                                  NaN
NaN                                     NaN
NaN                               NaN
NaN                           NaN                                  NaN
NaN                                            NaN
NaN                              NaN                              NaN
NaN                                  NaN
NaN                               NaN
NaN                                                NaN
NaN                                               NaN
NaN                                                     NaN
NaN                                            NaN
NaN                                      NaN
NaN                                       NaN
NaN                             NaN
NaN                             NaN
NaN                                    NaN
NaN                                       NaN
```

NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN NaN
NaN NaN
NaN NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN ZHAOHUIXIN Alwinner 1.8 GHz a13
1 B0D87RK5Q8 NaN
NaN #5,653 in Computers & Accessories (See Top 100 in Computers & Accessories) #670 in Traditional Laptop Computers 4.1 4.1 out of 5 stars \n 13 ratings \n\n\n 4.1 out of 5 stars
June 27, 2024 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN 1 Lithium Polymer batteries required. (included) TPV
Silver DDR3 SDRAM
128 MB NaN
NaN NaN
14.09 x 8.97 x 0.86 inches 5.15 pounds
NaN 1.0
Windows 11 Pro NaN
NaN NaN
Intel 14.09 x 8.97 x 0.86 inches
NaN AceBook 7.6 Volts
309.99 NaN TPV
Core i5 NaN NaN
Silver NaN
NaN NaN
Integrated NaN
NaN NaN 512 GB
NaN NaN
NaN NaN
NaN NaN AceBook

Windows 11 Pro NaN
NaN NaN 16 GB
NaN 15.6 Inches Webcam
NaN NaN
NaN 4.5 out of 5 stars NaN
5 Hours NaN Integrated
Intel NaN
NaN NaN Intel UHD
Graphics 617 512 GB SSD
NaN NaN NaN
1920x1080 Pixels NaN
NaN NaN
2.0 3.6 GHz core_i5 NaN
16 GB LPDDR3 1920 x 1080 pixels
15.6 Inches 802.11a/b/g/n/ac TPV 15.6" Laptop Computer (Intel Core i5 / 16GB RAM/ 512GB SSD), MS Office 2024, FHD Display with 100% sRGB Color Gamut, Windows 11 Pro Notebook PC with Dual Band Wi-Fi, Webcam (Silver) $369.99
https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzMzE1MDUzNjc5MDI6OjA6Og&url=%2FTPV-Computer-Display-Windows-Notebook%2Fdp%2FB0D87RK5Q8%2Fref%3Dsr_1_2_sspa%3Fdib%3DeyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-2-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN

```
NaN                                  NaN                                      NaN
NaN                                        NaN
NaN                                    NaN
NaN                                             NaN
NaN                                   NaN
NaN                                    NaN
NaN                                         NaN
NaN                                           NaN
NaN                                           NaN
NaN                                  NaN                                      NaN
NaN                                        NaN
NaN                                                                 NaN
NaN                                              NaN
NaN                          NaN                                      NaN
NaN                                      NaN
NaN                                    NaN
NaN                                       NaN
NaN                                  NaN
NaN                               NaN                                          NaN
NaN                                           NaN
NaN                                  NaN                                      NaN
NaN                                       NaN
NaN                                  NaN
NaN                                                  NaN
NaN                                                 NaN
NaN                                                        NaN
NaN                                              NaN
NaN                                         NaN
NaN                                          NaN
NaN                                  NaN
NaN                                  NaN
NaN                                        NaN
NaN                                           NaN
NaN                                        NaN
NaN                                   NaN
NaN                                                          NaN
NaN                                           NaN
NaN                                              NaN
NaN                                        NaN
NaN                               NaN
NaN                                        NaN                          NaN
NaN                                  NaN
NaN                                 NaN                                NaN
NaN                                  NaN
NaN                                       NaN
NaN                                              NaN
NaN                                   NaN
NaN          TPV            Intel          Core i5
```

```
# For OS
osColumnsFiltered = [os for os in columns if 'operating'.lower() in
os.lower() ]
osColumnsFiltered
```

```
['Other Technical Details_Operating System',
 'Product Details_Operating System',
 'New Product Details_Operating System']
```

```
osColumns = ['Other Technical Details_Operating System','Product
Details_Operating System','New Product Details_Operating System']
df = df.copy()
df['Operating_System'] = df[osColumns].bfill(axis=1).iloc[:,0]

# For RAM
RAMColumnsFiltered = [ram for ram in columns if 'ram'.lower() in
ram.lower() ]
RAMColumnsFiltered
```

```
['Additional Details_Graphics Card Ram Size',
 'Additional Details_Ram Memory Installed Size',
 'Product Details_RAM Memory Technology',
 'Product Details_RAM Type',
 'Product Details_Ram Memory Installed Size',
 'Technical Details_Graphics Card Ram Size',
 'Technical Details_RAM',
 'New Product Details_Graphics Ram Type',
 'New Product Details_RAM Memory Installed',
 'New Product Details_RAM Memory Slot Total Count',
 'New Product Details_RAM Memory Technology',
 'New Product Details_RAM Type',
 'New Product Details_Ram Memory Maximum Size',
 'New Product Details_Graphics Card Ram',
 'New Product Details_RAM']
```

```
ramColumns = ['Product Details_Ram Memory Installed Size','Product
Details_Memory Storage Capacity','New Product Details_Ram Memory
Maximum Size','New Product Details_RAM Memory Installed','Technical
Details_RAM','New Product Details_RAM']
df = df.copy()
df['RAM_Size'] = df[ramColumns].bfill(axis=1).iloc[:,0]

# For Storage
storageColumnsFiltered = [storage for storage in columns if
'storage'.lower() in storage.lower() or 'drive'.lower() in
storage.lower() or 'disk'.lower() in storage.lower() ]
storageColumnsFiltered
```

```
['Additional Details_Hard Drive Size',
 'Other Technical Details_Hard Drive Interface',
 'Other Technical Details_Hard Drive Rotational Speed',
```

```
 'Other Technical Details_Optical Drive Type',
 'Product Details_Hard Disk Description',
 'Product Details_Hard Disk Size',
 'Product Details_Memory Storage Capacity',
 'Technical Details_Hard Drive',
 'New Product Details_Hard Disk Description',
 'New Product Details_Hard Disk Interface',
 'New Product Details_Hard-Drive Size',
 'New Product Details_Number of Drivers',
 'New Product Details_Optical Storage Device',
 'New Product Details_Hard Disk Rotational Speed',
 'New Product Details_Hard Drive',
 'New Product Details_Hard Drive Interface',
 'New Product Details_Hard Drive Rotational Speed',
 'New Product Details_Memory Storage Capacity',
 'New Product Details_Optical Drive Type']
```

```
storageColumns = ['Technical Details_Hard Drive','Additional
Details_Hard Drive Size','Product Details_Hard Disk Size','New Product
Details_Hard-Drive Size','New Product Details_Memory Storage
Capacity']
df = df.copy()
df['Storage'] = df[storageColumns].bfill(axis=1).iloc[:,0]

# For Display
displayColumnsFiltered = [storage for storage in columns if
'display'.lower() in storage.lower() or 'screen'.lower() in
storage.lower() ]
displayColumnsFiltered
```

```
['Additional Details_Standing screen display size',
 'Product Details_Display Resolution Maximum',
 'Product Details_Display resolution',
 'Product Details_Screen Size',
 'Technical Details_Max Screen Resolution',
 'Technical Details_Screen Resolution',
 'Technical Details_Standing screen display size',
 'New Product Details_Display Resolution Maximum',
 'New Product Details_Display Technology',
 'New Product Details_Display Type',
 'New Product Details_Has Color Screen',
 'New Product Details_Screen Finish',
 'New Product Details_Screen Size',
 'New Product Details_Display Refresh Rate in Hertz',
 'New Product Details_Max Screen Resolution',
 'New Product Details_Maximum Display Brightness',
 'New Product Details_Screen Bezel Thickness',
 'New Product Details_Standing screen display size',
 'New Product Details_Touch Screen Type']
```

```
displaySizeColumns = ['Product Details_Screen Size','Technical
Details_Standing screen display size','New Product Details_Screen
Size','Additional Details_Standing screen display size']
df = df.copy()
df['Display_size'] = df[displaySizeColumns].bfill(axis=1).iloc[:,0]

# For Laptop Model Name
modelColumnsFiltered = [model for model in columns if 'model'.lower()
in model.lower()]
modelColumnsFiltered
```

```
['Additional Details_Item model number',
 'Other Technical Details_Item model number',
 'Product Details_CPU Model',
 'Product Details_Model Name',
 'Technical Details_Item model number',
 'New Product Details_CPU Model Generation',
 'New Product Details_CPU Model Number',
 'New Product Details_CPU Model Speed Maximum',
 'New Product Details_Model Name',
 'New Product Details_Model Number',
 'New Product Details_Model Year',
 'New Product Details_Item model number']
```

```
modelColumns = ['New Product Details_Model Name','New Product
Details_Item model number','Product Details_Model Name','New Product
Details_Model Name','Other Technical Details_Item model
number','Additional Details_Item model number']
df = df.copy()
df['Laptop_Model_Name'] = df[modelColumns].bfill(axis=1).iloc[:,0]

# For Laptop Reviews
reviewColumnsFiltered = [review for review in columns if
'review'.lower() in review.lower()]
reviewColumnsFiltered
```

```
['Additional Details_Customer Reviews', 'New Product Details_Customer
Reviews']
```

```
reviewColumns = ['Additional Details_Customer Reviews', 'New Product
Details_Customer Reviews']
df = df.copy()
df['Number_of_reviews'] = df[reviewColumns].bfill(axis=1).iloc[:,0]

# Extract review counts from the column
df = df.copy()
df['reviews_count'] = df['Number_of_reviews'].str.extract(r'(\d+)\
s+rating(?:s)?')
df['reviews_count'] = df['reviews_count'].fillna(0)
df['reviews_count'] = df['reviews_count'].astype(int)
df.head(2)
```

Additional Details_ASIN Additional Details_Batteries Additional Details_Batteries required
Additional Details_Best Sellers Rank
Additional Details_Customer Reviews Additional Details_Date First Available Additional Details_Form Factor Additional Details_Graphics Card Ram Size Additional Details_Hard Drive Size Additional Details_Included Components Additional Details_Is Discontinued By Manufacturer Additional Details_Item Weight Additional Details_Item model number Additional Details_Manufacturer Additional Details_Number of Ports Additional Details_Processor Speed Additional Details_Product Dimensions Additional Details_Ram Memory Installed Size Additional Details_Resolution Additional Details_Scanner Resolution Additional Details_Specific instructions for use Additional Details_Standing screen display size Additional Details_Total Usb Ports Additional Details_Warranty Description Other Technical Details_Audio-out Ports (#) Other Technical Details_Batteries Other Technical Details_Brand Other Technical Details_Color Other Technical Details_Computer Memory Type Other Technical Details_Flash Memory Size Other Technical Details_Hard Drive Interface Other Technical Details_Hard Drive Rotational Speed Other Technical Details_Hardware Platform Other Technical Details_Item Dimensions LxWxH Other Technical Details_Item Weight Other Technical Details_Item model number Other Technical Details_Number of Processors Other Technical Details_Operating System Other Technical Details_Optical Drive Type Other Technical Details_Package Dimensions Other Technical Details_Power Source Other Technical Details_Processor Brand Other Technical Details_Product Dimensions Other Technical Details_Rear Webcam Resolution Other Technical Details_Series Other Technical Details_Voltage Price Product Details_Battery Cell Composition Product Details_Brand Product Details_CPU Model Product Details_CPU Speed Product Details_Cache Size Product Details_Color Product Details_Connectivity Technology Product Details_Display Resolution Maximum Product Details_Display resolution Product Details_Graphics Card Description Product Details_Graphics Coprocessor Product Details_Graphics Processor Manufacturer Product Details_Hard Disk Description Product Details_Hard Disk Size Product Details_Has webcam capability? Product Details_Human Interface Input Product Details_Item Weight Product Details_Lithium Battery Energy Content Product Details_Manufacturer Product Details_Memory Slots Available Product Details_Memory Storage Capacity Product Details_Model Name Product Details_Operating System Product Details_Processor Count Product Details_RAM Memory Technology Product Details_RAM Type Product Details_Ram Memory Installed Size Product Details_Resolution Product Details_Screen Size Product Details_Special Feature Product Details_Specific Uses For Product Product Details_Total USB Ports Product Details_Wireless Communication Technology Rating Technical Details_ASIN Technical Details_Average Battery Life (in hours) Technical Details_Batteries Technical Details_Card Description Technical Details_Chipset Brand Technical Details_Country of Origin Technical Details_Date First Available Technical Details_Graphics Card

Ram Size Technical Details_Graphics Coprocessor Technical Details_Hard Drive Technical Details_Item Weight Technical Details_Item model number Technical Details_Manufacturer Technical Details_Max Screen Resolution Technical Details_Memory Speed Technical Details_National Stock Number Technical Details_Number of USB 2.0 Ports Technical Details_Number of USB 3.0 Ports Technical Details_Processor Technical Details_Product Dimensions Technical Details_RAM Technical Details_Screen Resolution Technical Details_Standing screen display size Technical Details_Wireless Type
Title Typical Price
URL New Product Details_ASIN New Product Details_Additional Features New Product Details_Audio Output Type New Product Details_Audio Recording New Product Details_Audio features New Product Details_Automatic Backup Software Included New Product Details_Battery Average Life New Product Details_Battery Average Life Standby New Product Details_Battery Cell Type New Product Details_Best Sellers Rank New Product Details_Bluetooth Version New Product Details_Bluetooth support? New Product Details_Brand Name New Product Details_CPU Model Generation New Product Details_CPU Model Number New Product Details_CPU Model Speed Maximum New Product Details_Camera Description New Product Details_Chipset Type New Product Details_Color New Product Details_Connectivity Technology New Product Details_Control Method New Product Details_Customer Reviews New Product Details_Display Resolution Maximum New Product Details_Display Technology New Product Details_Display Type New Product Details_Form Factor New Product Details_Graphics Coprocessor New Product Details_Graphics Description New Product Details_Graphics Ram Type New Product Details_Hard Disk Description New Product Details_Hard Disk Interface New Product Details_Hard-Drive Size New Product Details_Has Color Screen New Product Details_Human-Interface Input New Product Details_Included Components New Product Details_Item Dimensions L x W x Thickness New Product Details_Keyboard Description New Product Details_Keyboard Layout New Product Details_Manufacturer New Product Details_Microphone Form Factor New Product Details_Model Name New Product Details_Model Number New Product Details_Model Year New Product Details_Native Resolution New Product Details_Number of Drivers New Product Details_Number of Ethernet Ports New Product Details_Number of Ports New Product Details_Operating System New Product Details_Optical Storage Device New Product Details_Power Device New Product Details_Processor Brand New Product Details_Processor Count New Product Details_Processor Series New Product Details_Processor Speed New Product Details_RAM Memory Installed New Product Details_RAM Memory Slot Total Count New Product Details_RAM Memory Technology New Product Details_RAM Type New Product Details_Ram Memory Maximum Size New Product Details_Resolution New Product Details_Screen Finish New Product Details_Screen Size New Product Details_Series Number New Product Details_Speaker Description New Product Details_Specific Uses For Product New Product Details_Supported Monitor Maximum Quantity New Product Details_Total Number of HDMI Ports New Product Details_Total Thunderbolt Ports New

```
Product Details_Total Usb Ports  New Product Details_UPC New Product
Details_Video Output New Product Details_Video Processor New Product
Details_Virtual Reality Ready New Product Details_Webcam Capability
New Product Details_Wi-Fi Generation New Product Details_Wireless
Compability New Product Details_Wireless Technology New Product
Details_Age Range Description  New Product Details_Aspect Ratio  New
Product Details_Available M2 Slot Count New Product Details_Batteries
New Product Details_Battery Capacity New Product Details_Battery Power
New Product Details_Biometric Security Feature New Product
Details_Brand New Product Details_CPU Codename New Product Details_CPU
L3 Cache New Product Details_Cache Memory Installed Size New Product
Details_Card Description New Product Details_Cellular Technology New
Product Details_Chipset Brand New Product Details_Date First Available
New Product Details_Display Refresh Rate in Hertz New Product
Details_Flash Memory Size New Product Details_Front Photo Sensor
Resolution New Product Details_Generation  New Product Details_Global
Trade Identification Number New Product Details_Graphics Card Ram New
Product Details_Hard Disk Rotational Speed New Product Details_Hard
Drive New Product Details_Hard Drive Interface New Product
Details_Hard Drive Rotational Speed New Product Details_Hardware
Connectivity New Product Details_Hardware Interface New Product
Details_Is Electric New Product Details_Item Dimensions  LxWxH New
Product Details_Item Weight New Product Details_Item model number New
Product Details_LAN Port Bandwidth New Product Details_Lithium-Battery
Energy Content New Product Details_Max Screen Resolution New Product
Details_Maximum Display Brightness New Product Details_Memory Clock
Speed New Product Details_Memory Slots Available New Product
Details_Memory Speed New Product Details_Memory Storage Capacity New
Product Details_Notebook Pointing Device Description  New Product
Details_Number Of Cells  New Product Details_Number of Processors  New
Product Details_Number of Rear Facing Cameras  New Product
Details_Number of USB 2.0 Ports  New Product Details_Number of USB 3.0
Ports New Product Details_Optical Drive Type New Product Details_Photo
Sensor Resolution New Product Details_Processor New Product
Details_Processor Description New Product Details_Product Dimensions
New Product Details_RAM New Product Details_Rear Facing Camera Photo
Sensor Resolution New Product Details_Refresh Rate New Product
Details_Screen Bezel Thickness New Product Details_Sensor Type New
Product Details_Series New Product Details_Standing screen display
size New Product Details_Style Number  New Product Details_Total PCIe
Ports New Product Details_Touch Screen Type New Product
Details_Touchpad Feature New Product Details_Video Capture Resolution
New Product Details_Voltage New Product Details_Warranty Type New
Product Details_Wireless Communication Technology       Brand
Processor_Brand Processor_Model Operating_System RAM_Size     Storage
Display_size Laptop_Model_Name
Number_of_reviews  reviews_count
0              B0D8VSDCMK                                NaN
NaN  #81,691 in Computers & Accessories (See Top 100 in Computers &
Accessories)  #16,532 in Traditional Laptop Computers    5.0  5.0 out
```

of 5 stars \n 1 rating \n\n\n 5.0 out of 5 stars
July 5, 2024 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN 1 Lithium Polymer
batteries required. (included) ZHAOHUIXIN
Blue NaN
64 GB NaN
NaN NaN
9.53 x 6.89 x 0.83 inches 2.97 pounds
NaN 4.0
Android BD-R
NaN NaN
Alwinner 9.53 x 6.89 x 0.83 inches
NaN PC1068 NaN
119.99 NaN ZHAOHUIXIN
NaN NaN NaN
NaN NaN
1280x800 Pixels NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN 2 GB PC1068
NaN NaN
NaN NaN NaN
NaN 10.1 Inches NaN
NaN NaN
NaN 4.5 out of 5 stars NaN
NaN NaN Integrated
ARM NaN
NaN NaN
NaN 64 GB Emmc NaN
NaN NaN 1280x800
Pixels NaN
NaN 1.0
1.0 1.8 GHz a13 NaN
DDR4 NaN
10.1 Inches NaN
Mini Android 12 Laptop Computer, Portable Small Netbook with Allwinner A133 CPU Android 12 OS 2GB RAM 64GB EMMC HD IPS Screen 1920x800 0.3MP Camera (Blue) NaN https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZT

```
ozMDAzODc3MDQzMzg2MDI6OjA6Og&url=%2FZHAOHUIXIN-Computer-Portable-
Allwinner-1920x800%2Fdp%2FB0D8VSDCMK%2Fref%3Dsr_1_1_sspa%3Fdib
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-1-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1                      NaN
NaN                                   NaN
NaN                                NaN
NaN                                      NaN
NaN                                   NaN
NaN                                    NaN
NaN                            NaN
NaN                                  NaN
NaN                                    NaN
NaN                       NaN
NaN                                NaN
NaN                                            NaN
NaN                              NaN                               NaN
NaN                                      NaN
NaN                                       NaN
NaN                                 NaN
NaN                                       NaN
NaN                                                   NaN
NaN                                 NaN
NaN                                        NaN
NaN                              NaN                               NaN
NaN                                    NaN
NaN                                  NaN
NaN                                        NaN
NaN                                 NaN
NaN                                  NaN
NaN                                      NaN
NaN                                       NaN
NaN                                         NaN
NaN                               NaN                               NaN
NaN                                     NaN
NaN                                                     NaN
NaN                                          NaN
NaN                      NaN                              NaN
NaN                                       NaN
NaN                                  NaN
NaN                                     NaN
NaN                               NaN
NaN                           NaN                                    NaN
NaN                                            NaN
NaN                              NaN                                NaN
```

```
NaN                                  NaN
NaN                               NaN
NaN                                               NaN
NaN                                              NaN
NaN                                                    NaN
NaN                                           NaN
NaN                                     NaN
NaN                                      NaN
NaN                             NaN
NaN                             NaN
NaN                                    NaN
NaN                                       NaN
NaN                                    NaN
NaN                              NaN
NaN                                                     NaN
NaN                                       NaN
NaN                                          NaN
NaN                                    NaN
NaN                           NaN
NaN                                    NaN                     NaN
NaN                              NaN
NaN                             NaN                        NaN
NaN                              NaN
NaN                                   NaN
NaN                                          NaN
NaN                               NaN
NaN  ZHAOHUIXIN        Alwinner     1.8 GHz a13          Android     2
GB  64 GB Emmc  10.1 Inches            PC1068    5.0  5.0 out of 5
stars    \n    1 rating   \n\n\n 5.0 out of 5 stars              1
1              B0D87RK5Q8                          NaN
NaN      #5,653 in Computers & Accessories (See Top 100 in Computers &
Accessories)  #670 in Traditional Laptop Computers  4.1  4.1 out of 5
stars    \n    13 ratings   \n\n\n 4.1 out of 5 stars
June 27, 2024                           NaN
NaN                                NaN
NaN                                               NaN
NaN                                  NaN
NaN                                 NaN
NaN                                   NaN
NaN                           NaN
NaN                                             NaN
NaN                                 NaN
NaN                                          NaN  1 Lithium Polymer
batteries required. (included)                          TPV
Silver                                  DDR3 SDRAM
128 MB                                         NaN
NaN                                      NaN
14.09 x 8.97 x 0.86 inches                         5.15 pounds
NaN                                          1.0
Windows 11 Pro                                        NaN
```

NaN NaN
Intel 14.09 x 8.97 x 0.86 inches
NaN AceBook 7.6 Volts
309.99 NaN TPV
Core i5 NaN NaN
Silver NaN
NaN NaN
Integrated NaN
NaN NaN 512
GB NaN
NaN NaN
NaN NaN
NaN NaN AceBook
Windows 11 Pro NaN
NaN NaN 16 GB
NaN 15.6 Inches Webcam
NaN NaN
NaN 4.5 out of 5 stars NaN
5 Hours NaN Integrated
Intel NaN
NaN NaN Intel UHD
Graphics 617 512 GB SSD
NaN NaN NaN
1920x1080 Pixels NaN
NaN NaN
2.0 3.6 GHz core_i5 NaN
16 GB LPDDR3 1920 x 1080 pixels
15.6 Inches 802.11a/b/g/n/ac TPV 15.6" Laptop Computer (Intel Core i5 / 16GB RAM/ 512GB SSD), MS Office 2024, FHD Display with 100% sRGB Color Gamut, Windows 11 Pro Notebook PC with Dual Band Wi-Fi, Webcam (Silver) $369.99
https://www.amazon.com/sspa/click?ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZTozMDAzMzE1MDUzNjc5MDI6OjA6Og&url=%2FTPV-Computer-Display-Windows-Notebook%2Fdp%2FB0D87RK5Q8%2Fref%3Dsr_1_2_sspa%3Fdib%3DeyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-2-spons%26sp_csd%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1 NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN
NaN NaN

```
NaN                                    NaN
NaN                       NaN
NaN                                NaN
NaN                                            NaN
NaN                              NaN                             NaN
NaN                                      NaN
NaN                                       NaN
NaN                                 NaN
NaN                                       NaN
NaN                                                   NaN
NaN                                 NaN
NaN                                        NaN
NaN                              NaN                             NaN
NaN                                    NaN
NaN                                  NaN
NaN                                        NaN
NaN                                 NaN
NaN                                  NaN
NaN                                      NaN
NaN                                       NaN
NaN                                         NaN
NaN                               NaN                             NaN
NaN                                     NaN
NaN                                                     NaN
NaN                                          NaN
NaN                      NaN                              NaN
NaN                                       NaN
NaN                                  NaN
NaN                                     NaN
NaN                               NaN
NaN                           NaN                                  NaN
NaN                                            NaN
NaN                              NaN                              NaN
NaN                                  NaN
NaN                               NaN
NaN                                                NaN
NaN                                               NaN
NaN                                                     NaN
NaN                                            NaN
NaN                                      NaN
NaN                                       NaN
NaN                             NaN
NaN                             NaN
NaN                                    NaN
NaN                                       NaN
NaN                                    NaN
NaN                              NaN
NaN                                                      NaN
NaN                                       NaN
NaN                                          NaN
```

```
NaN                                          NaN
NaN                               NaN
NaN                                          NaN                          NaN
NaN                                    NaN
NaN                                   NaN                                NaN
NaN                                    NaN
NaN                                         NaN
NaN                                                 NaN
NaN                                    NaN
NaN          TPV           Intel          Core i5   Windows 11 Pro    16
GB  512 GB SSD  15.6 Inches           AceBook  4.1  4.1 out of 5 stars
\n    13 ratings   \n\n\n 4.1 out of 5 stars             13
```

Many Amazon products had imperfect data

Manually Checking for inconsistent data using domain knowledge

```
df['Processor_Brand'].unique()
```

```
array(['Alwinner', 'Intel', nan, 'AMD', 'MediaTek', 'Apple',
'Celeron',
       'Qualcomm', 'HP', 'I', 'Jasper Lake', 'core_i7_5650u', 'ARM',
       'Dell', 'Core', 'core_i7_6700hq', 'Intel Celeron N4120'],
      dtype=object)
```

```
# All these Processor Brands are misnomers, they are listed as
processor brands on Amazon but are instead Series or a range of
processors by Intel
df.loc[df['Processor_Brand'] == 'Celeron',['Processor_Model']] =
'Celeron N5095'
df.loc[df['Processor_Brand'] == 'Celeron',['Processor_Brand']] =
'Intel'

df.loc[df['Processor_Brand'] == 'I',['Processor_Brand']] = 'Intel'

df.loc[df['Processor_Brand'] == 'Jasper Lake',['Processor_Model']] =
'Celeron N5095'
df.loc[df['Processor_Brand'] == 'Jasper Lake',['Processor_Brand']] =
'Intel'

df.loc[df['Processor_Brand'] == 'Core',['Processor_Brand']] = 'Intel'

df.loc[df['Processor_Brand'] == 'Intel Celeron N4120',
['Processor_Model']] = df.loc[df['Processor_Brand'] == 'Intel Celeron
N4120',['Processor_Brand']]

df.loc[df['Processor_Brand'] == 'Intel Celeron N4120',
['Processor_Brand']] = 'Intel'

df.loc[df['Processor_Brand'] == 'HP',['Processor_Brand']] = 'Intel'
```

```
df.loc[df['Processor_Brand'] == 'Dell',['Processor_Brand']] = 'Intel'

df.loc[df['Processor_Brand'] == 'core_i7_6700hq',['Processor_Brand']]
= 'Intel'

df.loc[df['Processor_Brand'] == 'core_i7_5650u',['Processor_Brand']] =
'Intel'

# Some products have Processor name as 'Intel Core i5' at the cost of
not giving the processor comapny explicitly
df.loc[df['Processor_Model'].str.contains('Apple',na=False),'Processor
_Brand'] = 'Apple'
df.loc[df['Processor_Model'].str.startswith('Intel',na=False),'Process
or_Brand'] = 'Intel'
```

Only selecting common columns for analysis across data collected by other teammates across Flipkart and BestBuy

```
finalColumns =
['Brand','Laptop_Model_Name','Processor_Brand','Operating_System','Pro
cessor_Model','RAM_Size','Storage','Display_size','Rating','reviews_co
unt','Price','URL']
finalDF = df[finalColumns]
finalDF.head(2)
```

```
        Brand Laptop_Model_Name Processor_Brand Operating_System
Processor_Model RAM_Size     Storage Display_size              Rating
reviews_count   Price
URL
0  ZHAOHUIXIN            PC1068        Alwinner          Android
1.8 GHz a13     2 GB  64 GB Emmc  10.1 Inches  4.5 out of 5 stars
1  119.99  https://www.amazon.com/sspa/click?
ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZT
ozMDAzODc3MDQzMzg2MDI6OjA6Og&url=%2FZHAOHUIXIN-Computer-Portable-
Allwinner-1920x800%2Fdp%2FB0D8VSDCMK%2Fref%3Dsr_1_1_sspa%3Fdib
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-1-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1
1         TPV           AceBook           Intel   Windows 11 Pro
Core i5    16 GB  512 GB SSD  15.6 Inches  4.5 out of 5 stars
13  309.99            https://www.amazon.com/sspa/click?
ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZT
ozMDAzMzE1MDUzNjc5MDI6OjA6Og&url=%2FTPV-Computer-Display-Windows-
Notebook%2Fdp%2FB0D87RK5Q8%2Fref%3Dsr_1_2_sspa%3Fdib
```

```
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-2-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1

# Filtering out rows with exact same values(might have been introduced
due to multiple iterations of scraping)
# Many links scraped from Amazon are of affiliated products(sponsored)
which Amazon promotes across pagination - introducing duplicate rows.
duplicates = finalDF[finalDF.duplicated(keep=False)]
duplicates.shape

(55, 12)

# Drop duplicates and keep only the first occurrence
finalDF_duplicates_dropped = finalDF.drop_duplicates()
finalDF_duplicates_dropped.shape

(2123, 12)

# Columns to check for NaN values
columns_to_check_naN = ['Brand', 'Laptop_Model_Name',
'Processor_Brand', 'Operating_System',
       'Processor_Model', 'RAM_Size', 'Storage', 'Display_size',
'Price']

df_with_nan_in_specific_columns =
finalDF_duplicates_dropped[finalDF_duplicates_dropped[columns_to_check
_naN].isna().any(axis=1)]
df_with_nan_in_specific_columns.head(2)

    Brand Laptop_Model_Name Processor_Brand Operating_System
Processor_Model RAM_Size Storage Display_size              Rating
reviews_count  Price
URL
10  Apple       MacBook Air             NaN           Mac OS
Unknown     8 GB  256 GB  13.6 Inches  4.0 out of 5 stars
0  849.0                          https://www.amazon.com/2022-Apple-
MacBook-Laptop-chip/dp/B0B3BVWJ6Y/ref=sr_1_9?dib=eyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-
1DtQHxgmY&dib_tag=se&qid=1726997913&s=pc&sr=1-9
11  Apple       MacBook Pro           Apple           Mac OS    Apple
M3 Pro    18 GB  512 GB  14.2 Inches  4.7 out of 5 stars
```

```
0    NaN  https://www.amazon.com/Apple-MacBook-Laptop-11%E2%80%91core-14%E2%80%91core/dp/B0CM5JV26D/ref=sr_1_10?dib=eyJ2IjoiMSJ9.Mxv-LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vhpmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUubD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQYuN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY&dib_tag=se&qid=1726997913&s=pc&sr=1-10

print(f"Shape of rows to be dropped:
{df_with_nan_in_specific_columns.shape}")
finalDF_cleaned = finalDF_duplicates_dropped.dropna(subset =
columns_to_check_naN)
print(f"Shape after dropping : {finalDF_cleaned.shape}")

Shape of rows to be dropped: (125, 12)
Shape after dropping : (1998, 12)
```

Updating the DF before exporting to ensure collaboration with other teammates' data

```
# Ensuring data is formatted in a uniform structure.
finalDF_cleaned = finalDF_cleaned.copy()
finalDF_cleaned['Storage_Type'] = finalDF_cleaned['Storage'].apply(
    lambda x: 'HDD' if pd.Series(x).str.contains('HDD',
case=False).any() else
              'SSD' if pd.Series(x).str.contains('SSD',
case=False).any() else
              'EMMC' if pd.Series(x).str.contains('EMMC',
case=False).any() else 'SSD'
)

# Extract storage size into a new column 'Storage_Size'
finalDF_cleaned = finalDF_cleaned.copy()
finalDF_cleaned['Storage_Size'] =
finalDF_cleaned['Storage'].str.extract(r'(\d+\s*[KMG]B|\d+\s*TB|\d+\
s*SSD)', expand=False)

print(f"Shape of storage_size naN values :
{finalDF_cleaned[finalDF_cleaned['Storage_Size'].isna()].shape}")
finalDF_cleaned = finalDF_cleaned.dropna(subset=['Storage_Size'])
print(f"Shape after removing rows which had missing storage sizes :
{finalDF_cleaned.shape}")

Shape of storage_size naN values : (68, 14)
Shape after removing rows which had missing storage sizes : (1930, 14)

# Converting all storage sizes to GB
def convert_to_gb(size):
    if 'TB' in size:
        return int(size.replace('TB', '').strip()) * 2048
```

```
    elif 'GB' in size:
        return int(size.replace('GB', '').strip())
    elif 'SSD' in size:
        return int(size.replace('SSD', '').strip())
    else:
        return size

finalDF_cleaned = finalDF_cleaned.copy()
finalDF_cleaned['Storage_Size_GB'] =
finalDF_cleaned['Storage_Size'].apply(convert_to_gb)
finalDF_cleaned['Storage_Size_GB'] =
finalDF_cleaned['Storage_Size_GB'].astype(int)

# Extract the rating using REGEX
finalDF_cleaned = finalDF_cleaned.copy()
finalDF_cleaned['Extracted_Rating'] =
finalDF_cleaned['Rating'].str.extract(r'(\d+\.\d+|\d+)', expand=False)
finalDF_cleaned['Extracted_Rating'] =
finalDF_cleaned['Extracted_Rating'].astype(float)
finalDF_cleaned.head(2)
```

```
        Brand Laptop_Model_Name Processor_Brand Operating_System
Processor_Model RAM_Size     Storage Display_size              Rating
reviews_count   Price
URL Storage_Type Storage_Size  Storage_Size_GB  Extracted_Rating
0  ZHAOHUIXIN            PC1068        Alwinner          Android
1.8 GHz a13     2 GB  64 GB Emmc  10.1 Inches  4.5 out of 5 stars
1  119.99  https://www.amazon.com/sspa/click?
ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZT
ozMDAzODc3MDQzMzg2MDI6OjA6Og&url=%2FZHAOHUIXIN-Computer-Portable-
Allwinner-1920x800%2Fdp%2FB0D8VSDCMK%2Fref%3Dsr_1_1_sspa%3Fdib
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-1-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1         EMMC        64
GB              64              4.5
1         TPV          AceBook          Intel   Windows 11 Pro
Core i5    16 GB  512 GB SSD  15.6 Inches  4.5 out of 5 stars
13  309.99           https://www.amazon.com/sspa/click?
ie=UTF8&spc=MToyNTU2ODExODc2NTY2MDA3OjE3MjY5OTc5MTM6c3BfYXRmX2Jyb3dzZT
ozMDAzMzE1MDUzNjc5MDI6OjA6Og&url=%2FTPV-Computer-Display-Windows-
Notebook%2Fdp%2FB0D87RK5Q8%2Fref%3Dsr_1_2_sspa%3Fdib
%3DeyJ2IjoiMSJ9.Mxv-
LfaT1mRTkqi6GWEFXxFggO64cMc5a5WQAxAoDYKDc12AZYR8P_ulvGvs2fWDJ7_Nm3Q_vh
pmjYCsv0OJPJs6Bo1FRX66cFxFfjDS5M6onhimzcAeCOZ3ganbR1ztxCB3tNO3H2yyijUu
bD6xTB3G5UxB2MqPQQaHrdLyLai29xSPy1hZkKf5Sm2MjOm9tSgk53w2mGq_T8vokhTRQY
```

```
uN1uCwbBymaj5IEXp_6tzKZ-DZ8lOcjCmWHFWVn2fkST3y58q3_y3AxTbeKQumI-
hyzZmMa7tonKGyaYVia00.eFMZpiqa4BbfF8ul13G1oFWFR-jOJGFy_-1DtQHxgmY
%26dib_tag%3Dse%26qid%3D1726997913%26s%3Dpc%26sr%3D1-2-spons%26sp_csd
%3Dd2lkZ2V0TmFtZT1zcF9hdGZfYnJvd3Nl%26psc%3D1          SSD       512
GB              512               4.5

# Remove 'inches' and keep only the number from Display Screen Size
finalDF_cleaned['Display_size_num'] =
finalDF_cleaned['Display_size'].str.replace(r'\s*inches?', '',
case=False, regex=True)

finalDF_cleaned['Display_size_num'] =
finalDF_cleaned['Display_size_num'].astype(float)

# Remove units from RAM Size ( all are in GB)
finalDF_cleaned['RAM'] = finalDF_cleaned['RAM_Size'].str.extract(r'(\
d+)').astype(int)

# Drop all columns we just extracted data from and drop redundant
columns
columns_to_drop =
['URL','Rating','Storage_Size','Storage','Display_size','RAM_Size']
finalDF_cleaned_aligned =
finalDF_cleaned.drop(columns=columns_to_drop)
finalDF_cleaned_aligned.head()

        Brand Laptop_Model_Name Processor_Brand Operating_System
Processor_Model  reviews_count    Price Storage_Type  Storage_Size_GB
Extracted_Rating  Display_size_num  RAM
0  ZHAOHUIXIN            PC1068        Alwinner          Android
1.8 GHz a13              1   119.99         EMMC               64
4.5              10.1    2
1         TPV           AceBook           Intel   Windows 11 Pro
Core i5             13   309.99          SSD              512
4.5              15.6   16
2          HP         Elitebook           Intel   Windows 11 Pro
Intel Core i7              5  1079.00         SSD             2048
4.0              16.0   32
3       Apple       MacBook Air           Apple           Mac OS
Apple M3              0   929.00          SSD               256
4.0              13.6    8
4       Apple       MacBook Air           Apple           Mac OS
Apple M3              0  1449.00          SSD               512
4.0              15.3   16
```

Renaming columns to ensure uniformity across teammates

```
# Renaming columns to ensure uniformity
renaming_dict = {
    'Brand':'Laptop_Brand',
    'Laptop_Model_Name':'Laptop_Name',
```

```
    'Processor_Brand':'Processor_Company',
    'Processor_Model':'Processor',
    'Storage_Size_GB':'Storage',
    'Display_size_num':'Screen_Size',
    'Extracted_Rating':'Rating',
    'reviews_count':'Number_of_Reviews'
}
finalDF_cleaned_aligned =
finalDF_cleaned_aligned.rename(columns=renaming_dict)

# Adding a Source column to identify the source of row datasets are
merged (Amazon, Flipkart or BestBuy)
finalDF_cleaned_aligned = finalDF_cleaned_aligned.copy()
finalDF_cleaned_aligned['Source'] = 'Amazon'
finalDF_cleaned_aligned.head()
```

```
  Laptop_Brand  Laptop_Name Processor_Company Operating_System
Processor  Number_of_Reviews    Price Storage_Type  Storage  Rating
Screen_Size  RAM  Source
0   ZHAOHUIXIN       PC1068          Alwinner          Android    1.8
GHz a13                  1   119.99         EMMC       64     4.5
10.1    2  Amazon
1          TPV      AceBook             Intel   Windows 11 Pro
Core i5                 13   309.99          SSD      512     4.5
15.6   16  Amazon
2           HP    Elitebook             Intel   Windows 11 Pro  Intel
Core i7                  5  1079.00          SSD     2048     4.0
16.0   32  Amazon
3        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0   929.00          SSD      256     4.0
13.6    8  Amazon
4        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0  1449.00          SSD      512     4.0
15.3   16  Amazon
```

```
print(f"Final Shape of cleaned data :
{finalDF_cleaned_aligned.shape}")
```

```
Final Shape of cleaned data : (1930, 13)
```

```
finalDF_cleaned_aligned.describe().T
```

```
                     count         mean           std    min     25%
50%        75%        max
Number_of_Reviews  1930.0    85.392746    176.604393   0.00    2.00
12.0    62.0000      996.0
Price              1930.0   698.844674    581.544264  44.79  299.99
569.0   908.7425     4939.0
Storage            1930.0  2563.029016  33743.047376   2.00  256.00
512.0  2048.0000  1048576.0
Rating             1928.0     4.353268      0.295933   1.00    4.30
```

```
4.4     4.5000        5.0
Screen_Size         1930.0    14.786067      1.676679   7.00   14.00
15.6    15.6000       18.0
RAM                 1930.0    21.147150     23.583480   2.00    8.00
16.0    32.0000      512.0
```

```
finalDF_cleaned_aligned.describe(include='object').T
```

```
                   count unique             top  freq
Laptop_Brand        1930     71              HP   430
Laptop_Name         1930    767        Latitude   152
Processor_Company   1930      7           Intel  1535
Operating_System    1930     45  Windows 11 Pro   748
Processor           1930    157         Core i5   190
Storage_Type        1930      3             SSD  1723
Source              1930      1          Amazon  1930
```

```
outputDataFilePath = './consolidated_amazon_laptop_data.csv'
finalDF_cleaned_aligned.to_csv(outputDataFilePath, index=False)
print(f"Data saved to : {outputDataFilePath}")
```

```
Data saved to : ./consolidated_amazon_laptop_data.csv
```

Combining all Data

```
amazonDataSetPath = r'./amazon/consolidated_amazon_laptop_data.csv'
flipkartDatasetPath = r'./flipkart_data/flipkart_laptop_cleaned.csv'
bestbuyDatasetPath = r'./bestBuy/laptops_data_Best_Buy_22_09_24.csv'

amazonDF = pd.read_csv(amazonDataSetPath)
flipkartDF = pd.read_csv(flipkartDatasetPath)
bestbuyDF = pd.read_csv(bestbuyDatasetPath)

# Bestbuy

# Renaming columns to ensure uniformity
renaming_dict_BB = {
    'Brand':'Laptop_Brand',
    'Laptop_name':'Laptop_Name',
    'Processor_Company_Name':'Processor_Company',
    'Processor_Model':'Processor',
    'Laptop_Storage_GB':'Storage',
    'Rating_5':'Rating',
    'No_Of_Reviews':'Number_of_Reviews' ,
    'Laptop_Price':'Price',
    'Laptop_Memory_GB':'RAM',
    'Laptop_Display_Size_in':'Screen_Size'
}
```

```
bestbuyDF_aligned = bestbuyDF.rename(columns=renaming_dict_BB)
bestbuyDF_aligned.head(2)
```

```
     Price Laptop_Brand Laptop_Colour  RAM Storage Screen_Size
Laptop_Name Laptop_Graphics  Number_of_Reviews Rating
Processor_Company Storage_Type Operating_System
Processor
0  $529.99           HP        Silver   16     512        15.6    No
Model     No Graphics              1939    4.6             Intel
SSD        Windows11            Core i7 - 15-DY5073DX
1  $749.99         Dell          Blue   16    1024          16
Inspiron     No Graphics                211    4.7              Intel
SSD        Windows11   Core Ultra - i7640-7366BLU-PUS
```

```
# Common Columns for combined dataset
columns_for_alignment = ['Laptop_Brand', 'Laptop_Name',
'Processor_Company', 'Operating_System',
       'Processor', 'Number_of_Reviews', 'Price', 'Storage_Type',
'Storage',
       'Rating', 'Screen_Size', 'RAM']

# Filtering out common columns
bestbuyDF_aligned_final = bestbuyDF_aligned[columns_for_alignment]
bestbuyDF_aligned_final.shape
```

```
(1284, 12)
```

```
# Adding Source Column to identify datapoint source
bestbuyDF_aligned_final = bestbuyDF_aligned_final.copy()
bestbuyDF_aligned_final['Source'] = 'BestBuy'
bestbuyDF_aligned_final.shape
```

```
(1284, 13)
```

```
bestbuyDF_aligned_final.head()
```

```
  Laptop_Brand Laptop_Name Processor_Company Operating_System
Processor  Number_of_Reviews    Price Storage_Type Storage Rating
Screen_Size  RAM   Source
0           HP    No Model             Intel        Windows11
Core i7 - 15-DY5073DX               1939  $529.99          SSD     512
4.6        15.6   16  BestBuy
1         Dell    Inspiron             Intel        Windows11   Core
Ultra - i7640-7366BLU-PUS                 211  $749.99          SSD
1024    4.7          16   16  BestBuy
2       Lenovo        Flex             Intel        Windows11
Core i3 - 82R700L4US                181  $329.99          SSD     256
4.5          14    8  BestBuy
3       Lenovo        Yoga          AMD Ryzen        Windows11
7 8840HS - 83DM0003US                  72  $649.99          SSD    1024
4.7          16   16  BestBuy
```

```
4         Dell    Inspiron             Intel        Windows11
Core i5 - i3520-5124BLK-PUS               604  $629.99          SSD
512    4.5          14    8  BestBuy

#Flipkart
flipkartDF = flipkartDF.copy()
flipkartDF['Source'] = 'Flipkart'
flipkartDF.head()

  Laptop_Brand        Laptop_Name Processor_Company
Processor Operating_System  RAM  Storage Storage_Type  Screen_Size
Rating  Number_of_Reviews     Price    Source
0         ASUS        Vivobook 15             Intel    Core i3 12th
Gen 1215U       Windows 11    8      512          SSD         12.0
4.2              360.0   35990.0  Flipkart
1         ASUS  ROG Strix Scar 16             Intel  Core i9 14th Gen
14900HX       Windows 11   32     2048          SSD         14.0
0.0                0.0  339990.0  Flipkart
2         ASUS     TUF Gaming F15             Intel   Core i5 12th Gen
12500H       Windows 11   16      512          SSD         12.0
4.5               43.0   75990.0  Flipkart
3         ASUS     TUF Gaming F17             Intel   Core i5 11th Gen
11400H       Windows 11   16      512          SSD         11.0
4.3              467.0   50990.0  Flipkart
4         ASUS        Vivobook 15             Intel    Core i3 12th
Gen 1215U       Windows 11    8      512          SSD         12.0
4.2              360.0   35990.0  Flipkart

# Normalizing the price from INR to USD @ 83.97 INR = 1 USD [Exchange
Rate As of 07 October 2024]
conversion_rate = 83.97
flipkartDF['Price'] = round(flipkartDF['Price']/83.97,2)
flipkartDF.head()

  Laptop_Brand        Laptop_Name Processor_Company
Processor Operating_System  RAM  Storage Storage_Type  Screen_Size
Rating  Number_of_Reviews    Price    Source
0         ASUS        Vivobook 15             Intel    Core i3 12th
Gen 1215U       Windows 11    8      512          SSD         12.0
4.2              360.0   428.61  Flipkart
1         ASUS  ROG Strix Scar 16             Intel  Core i9 14th Gen
14900HX       Windows 11   32     2048          SSD         14.0
0.0                0.0  4048.95  Flipkart
2         ASUS     TUF Gaming F15             Intel   Core i5 12th Gen
12500H       Windows 11   16      512          SSD         12.0
4.5               43.0   904.97  Flipkart
3         ASUS     TUF Gaming F17             Intel   Core i5 11th Gen
11400H       Windows 11   16      512          SSD         11.0
4.3              467.0   607.24  Flipkart
4         ASUS        Vivobook 15             Intel    Core i3 12th
```

```
Gen 1215U       Windows 11    8      512          SSD         12.0
4.2              360.0   428.61  Flipkart
```

```
amazonDF.head()
```

```
  Laptop_Brand  Laptop_Name Processor_Company Operating_System
Processor  Number_of_Reviews    Price Storage_Type  Storage  Rating
Screen_Size  RAM  Source
0   ZHAOHUIXIN       PC1068          Alwinner          Android    1.8
GHz a13                  1   119.99         EMMC       64     4.5
10.1    2  Amazon
1          TPV      AceBook             Intel   Windows 11 Pro
Core i5                 13   309.99          SSD      512     4.5
15.6   16  Amazon
2           HP    Elitebook             Intel   Windows 11 Pro  Intel
Core i7                  5  1079.00          SSD     2048     4.0
16.0   32  Amazon
3        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0   929.00          SSD      256     4.0
13.6    8  Amazon
4        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0  1449.00          SSD      512     4.0
15.3   16  Amazon
```

```
# Combining all the datasets
master_df = pd.concat([amazonDF, flipkartDF,bestbuyDF_aligned_final],
ignore_index=True, sort=False)
master_df.head()
```

```
  Laptop_Brand  Laptop_Name Processor_Company Operating_System
Processor  Number_of_Reviews   Price Storage_Type Storage Rating
Screen_Size  RAM  Source
0   ZHAOHUIXIN       PC1068          Alwinner          Android    1.8
GHz a13                1.0  119.99         EMMC      64    4.5
10.1    2  Amazon
1          TPV      AceBook             Intel   Windows 11 Pro
Core i5               13.0  309.99          SSD     512    4.5
15.6   16  Amazon
2           HP    Elitebook             Intel   Windows 11 Pro  Intel
Core i7                5.0  1079.0          SSD    2048    4.0
16.0   32  Amazon
3        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0.0   929.0          SSD     256    4.0
13.6    8  Amazon
4        Apple  MacBook Air             Apple           Mac OS
Apple M3                 0.0  1449.0          SSD     512    4.0
15.3   16  Amazon
```

```
master_df.shape
```

```
(4838, 13)
```

Exporting The Aggregated Data

```
combinedOutputDataFilePath = './data/laptrack.csv'
master_df.to_csv(combinedOutputDataFilePath, index=False)
print(f"Data saved to : {combinedOutputDataFilePath}")
```

```
Data saved to : ./data/laptrack.csv
```

References

1. HTML Error Codes: Used for refrencing response codes and their meaning.
2. Request Module of Python: Used for creating custom headers during scraping
3. Intel Processor Products: Used to improve domain knowledge of processors series and ranges.
4. Regex 101: Used to debug the regex patterns.
5. Regex Part 1 By Real Python: Used for better clarity on certain symbols and patterns in regex.