CODERSCAVE ANALYST INTERN

Domain : Buisness Analytics Intern

Normal Task

Data Cleaning and Preprocessing

Problem Statement:=

1.Data Cleaning and Preprocessing The Objective of this project is to perform Data cleaning and Preprocessing of the provided Dataset to systematically clean, transform, and prepare the data to make it suitable for analysis, modeling, and decision-making. The aim of this project is to ensure that data integrity is maintained, and high- quality, reliable data is made available for informed and accurate business insights, thereby enhancing the effectiveness of data-driven decision-making processes.(Use Amazon's Dataset)

Introduction

This dataset contains more than a thousand rows of product ratings and reviews listed on Amazon platform by their users.

The variables of the data are;

product_id - Product ID

product_name - Name of the Product

category - Category of the Product

discounted_price - Discounted Price of the Product

actual_price - Actual Price of the Product

discount_percentage - Percentage of Discount for the Product

rating - Rating of the Product

rating_count - Number of user who contributed to the Amazon rating

about_product - Description about the Product

user_id - ID of the user who wrote review for the Product

user_name - Name of the user who wrote review for the Product

review_id - ID of the user review

review_title - Short review

review_content - Long review

img_link - Image Link of the Product

product_link - Official Website Link of the Product

The objective of this work is on data cleansing.

In [1]:
```
# Import required libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns

# Load raw data
data = pd.read_csv(r"C:\Users\HP\Desktop\Intern\CodersCave Bussiness Analytics Intern\Phase 1- Task 1\amazon.csv")
```

In [2]:
```
# View properties of dataset
data.info()

# Data type ‘object‘ represents strings, hence there is a need to convert data type accordingly
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1465 entries, 0 to 1464
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype 
---  ------               --------------  ----- 
 0   product_id           1465 non-null   object
 1   product_name         1465 non-null   object
 2   category             1465 non-null   object
 3   discounted_price     1465 non-null   object
 4   actual_price         1465 non-null   object
 5   discount_percentage  1465 non-null   object
 6   rating               1465 non-null   object
 7   rating_count         1463 non-null   object
 8   about_product        1465 non-null   object
 9   user_id              1465 non-null   object
 10  user_name            1465 non-null   object
 11  review_id            1465 non-null   object
 12  review_title         1465 non-null   object
 13  review_content       1465 non-null   object
 14  img_link             1465 non-null   object
 15  product_link         1465 non-null   object
dtypes: object(16)
memory usage: 183.3+ KB
```

In [3]:
```
# List total number of rows and columns
print("This dataset contains ", data.shape[0], " rows and ", data.shape[1], " columns")
```

```
This dataset contains  1465  rows and  16  columns
```

In [4]:
```
# Clean data header name
namechange={"product_id":"Product ID", "product_name":"Product Name", "category":"Category", "discounted_price":"Discou
data = data.rename(columns=namechange)
```

In [5]:
```
# Understand first 5 rows of data
data.head()
```

Out[5]:

	Product ID	Product Name	Category	Discounted Price	Actual Price	Discount Percentage	Rating	Rating Count	About Product	
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	₹399	₹1,099	64%	4.2	24,269	High Compatibility : Compatible With iPhone 12...	A
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹349	43%	4.0	43,994	Compatible with all Type C enabled devices, be...	
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	₹199	₹1,899	90%	3.9	7,928	【 Fast Charger& Data Sync】- With built-in safet...	A
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	₹329	₹699	53%	4.2	94,363	The boAt Deuce USB 300 2 in 1 cable is compati...	
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	₹154	₹399	61%	4.2	16,905	[CHARGE & SYNC FUNCTION]- This cable comes wit...	

In [6]:

```
# Check for duplicated rows
if data.duplicated().sum() == 0:
    print("No change in dataset as there is no duplicated rows.")
    print("Dataset remains as ", data.shape[0], " rows and ", data.shape[1], " columns")
else:
    # in the event of duplicated row, row with latest built year remains
    data = data.sort_values('YearBuilt').drop_duplicates(keep='last')
    print("Number of duplicated rows removed: ", data.duplicated().sum())
    print("After removing duplicated rows, dataset contains ", data.shape[0], " rows and ", data.shape[1], " columns")
```

```
No change in dataset as there is no duplicated rows.
Dataset remains as  1465  rows and  16  columns
```

In [7]:
```
# Check for null value in dataset
nullvalue = data.isna().sum()
nullvalue = nullvalue[nullvalue>0] / len(data) * 100
nullvalue = nullvalue.round(2).to_frame('%Null').sort_values('%Null', axis=0, ascending=False)
nullvalue
```

Out[7]:

	%Null
Rating Count	0.14

In [8]:
```
# Find out isnull rows in dataset
data[data["Rating Count"].isnull()] # 2 out of 1465 (0.14%) contains null
data = data.loc[~data['Rating Count'].isnull()] # remove from dataset
```

In [9]:
```
# All row for Discount Price column is in Indian Rupee '₹'
data['Discounted Price'].str.contains('₹').sum() # 1463 out of 1463
data['Discounted Price'].str.contains('.').sum() # 1463 out of 1463
```

Out[9]:
```
1463
```

In [10]:
```
# All row for Actual Price column is in Indian Rupee '₹'
data['Actual Price'].str.contains('₹').sum() # 1463 out of 1463
data['Actual Price'].str.contains('.').sum() # 1463 out of 1463
```

Out[10]:
```
1463
```

In [11]:
```
# All row for percentage column contains '%'
data['Discount Percentage'].str.contains('%').sum() # 1463 out of 1463
data['Discount Percentage'].str.contains('.').sum() # 1463 out of 1463
```

Out[11]:
```
1463
```

In [12]:
```
# Majority of rating count column contains ','
data['Rating Count'].str.contains(',').sum() # 1137 out of 1463
rc = data[~data['Rating Count'].str.contains(',', na=False)]
rc['Rating Count'].astype('float64').max() # remaining 326 out of 1463 is due to rating count lower than 1,000
```

Out[12]:
```
992.0
```

In [13]:
```
# Replace symbols for relevant columns
data['Discounted Price'] = data['Discounted Price'].str.replace("₹",'').str.replace(",",'')
data['Actual Price'] = data['Actual Price'].str.replace("₹",'').str.replace(",",'')
data['Discount Percentage'] = data['Discount Percentage'].str.replace("%",'')
data['Rating Count'] = data['Rating Count'].str.replace(",",'')
```

In [14]:
```
# Convert data type to float
data = data.astype({'Discounted Price': 'float64', 'Actual Price': 'float64', 'Discount Percentage': 'float64', 'Rating

# Error occurred when converting rating. (ValueError: could not convert string to float: '|')

#Other alternatives to covert data type
#data['discounted_price'] = data['discounted_price'].astype('float64')
#data['actual_price'] = pd.to_numeric(data['actual_price'])
```

In [15]:
```
# Resolve error for the conversion of rating
data.query('Rating == "|"') # 1 row without inputs for rating column.
data = data.loc[data['Rating'] != '|'] # remove column from dataset

# Convet data type to float
data['Rating'] = data['Rating'].astype('float64')
```

In [16]:
```
#Updates on dataset
data.info()
data.head()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1462 entries, 0 to 1464
Data columns (total 16 columns):
 #   Column               Non-Null Count  Dtype  
---  ------               --------------  -----  
 0   Product ID           1462 non-null   object 
 1   Product Name         1462 non-null   object 
 2   Category             1462 non-null   object 
 3   Discounted Price     1462 non-null   float64
 4   Actual Price         1462 non-null   float64
 5   Discount Percentage  1462 non-null   float64
 6   Rating               1462 non-null   float64
 7   Rating Count         1462 non-null   float64
 8   About Product        1462 non-null   object 
 9   User ID              1462 non-null   object 
 10  User Name            1462 non-null   object 
 11  Review ID            1462 non-null   object 
 12  Review Title         1462 non-null   object 
 13  Review Content       1462 non-null   object 
 14  IMG Link             1462 non-null   object 
 15  Product Link         1462 non-null   object 
dtypes: float64(5), object(11)
memory usage: 194.2+ KB
```

Out[16]:

	Product ID	Product Name	Category	Discounted Price	Actual Price	Discount Percentage	Rating	Rating Count	About Product
0	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	399.0	1099.0	64.0	4.2	24269.0	High Compatibility : Compatible With iPhone 12...
1	B098NS6PVG	Ambrane Unbreakable 60W / 3A Fast Charging 1.5...	Computers&Accessories|Accessories&Peripherals|...	199.0	349.0	43.0	4.0	43994.0	Compatible with all Type C enabled devices, be...
2	B096MSW6CT	Sounce Fast Phone Charging Cable & Data Sync U...	Computers&Accessories|Accessories&Peripherals|...	199.0	1899.0	90.0	3.9	7928.0	【 Fast Charger& Data Sync】- With built-in safet...
3	B08HDJ86NZ	boAt Deuce USB 300 2 in 1 Type-C & Micro USB S...	Computers&Accessories|Accessories&Peripherals|...	329.0	699.0	53.0	4.2	94363.0	The boAt Deuce USB 300 2 in 1 cable is compati...
4	B08CF3B7N1	Portronics Konnect L 1.2M Fast Charging 3A 8 P...	Computers&Accessories|Accessories&Peripherals|...	154.0	399.0	61.0	4.2	16905.0	[CHARGE & SYNC FUNCTION]- This cable comes wit...

In [17]:

```
# Explore split categories for ease of identification
catbreak = data['Category'].str.split('|', expand=True)
catbreak # Categories in the dataset can be broken down to 7 categories (0 to 6)

catbreaknull = catbreak.isna().sum()
catbreaknull = catbreaknull[catbreaknull>0] / len(catbreak) * 100
catbreaknull = catbreaknull.round(2).to_frame('%Null').sort_values('%Null', axis=0, ascending=False)
catbreaknull # to keep category 0 and 1

#catbreak[catbreak[2].isnull()]
```

Out[17]:

	%Null
6	99.11
5	94.25
4	64.50
3	11.29
2	0.55

In [18]:
```
# Split Categories, maintaining Category 0 (Product Division) and Category 1 (Product Group)
data[['Product Division', 'Product Group', 'Product Sub Group', 'Product Type', 'Gategory 4', 'Category 5', 'Category 6
```

In [19]:
```
# Explore breakdown of Product Division
data['Product Division'].value_counts()
#alternatively
#data[['Product Division']].groupby(['Product Division']).size()
#data['Product Division'].unique()
```

Out[19]:
```
Electronics              526
Computers&Accessories    451
Home&Kitchen             447
OfficeProducts            31
MusicalInstruments         2
HomeImprovement            2
Toys&Games                 1
Car&Motorbike              1
Health&PersonalCare        1
Name: Product Division, dtype: int64
```

In [20]:
```
# Explore breakdown of Product Group
data['Product Group'].value_counts()
```

Out[20]:
```
Accessories&Peripherals                    379
Kitchen&HomeAppliances                     307
HomeTheater,TV&Video                       162
Mobiles&Accessories                        161
Heating,Cooling&AirQuality                 116
WearableTechnology                          76
Headphones,Earbuds&Accessories              66
NetworkingDevices                           34
OfficePaperProducts                         27
ExternalDevices&DataStorage                 18
Cameras&Photography                         16
HomeStorage&Organization                    16
HomeAudio                                   16
GeneralPurposeBatteries&BatteryChargers     14
Accessories                                 14
Printers,Inks&Accessories                   11
CraftMaterials                               7
Components                                   5
OfficeElectronics                            4
Electrical                                   2
Monitors                                     2
Microphones                                  2
Arts&Crafts                                  1
PowerAccessories                             1
Tablets                                      1
Laptops                                      1
Kitchen&Dining                               1
CarAccessories                               1
HomeMedicalSupplies&Equipment                1
Name: Product Group, dtype: int64
```

In [21]:
```
# Clean data for Product Division and Product Group
data['Product Division'] = data['Product Division'].str.replace('&',' & ')
data['Product Division'] = data['Product Division'].str.replace('Products',' Products')
data['Product Division'] = data['Product Division'].str.replace('Instruments',' Instruments')
data['Product Division'] = data['Product Division'].str.replace('Improvement',' Improvement')
data['Product Division'] = data['Product Division'].str.replace('Personal',"Personal ")
data['Product Group'] = data['Product Group'].str.replace('&',' & ').str.replace(',',', ')
data['Product Group'] = data['Product Group'].str.replace('Appliances',' Appliances')
data['Product Group'] = data['Product Group'].str.replace('Theater',' Theater')
data['Product Group'] = data['Product Group'].str.replace('Quality',' Quality')
data['Product Group'] = data['Product Group'].str.replace('Technology',' Technology')
data['Product Group'] = data['Product Group'].str.replace('Devices',' Devices')
data['Product Group'] = data['Product Group'].str.replace('Paper',' Paper ')
data['Product Group'] = data['Product Group'].str.replace('Storage', ' Storage')
```

```
data['Product Group'] = data['Product Group'].str.replace('Purpose', ' Purpose ')
data['Product Group'] = data['Product Group'].str.replace('Chargers', ' Chargers')
data['Product Group'] = data['Product Group'].str.replace('Materials', ' Materials')
data['Product Group'] = data['Product Group'].str.replace('Electronics', ' Electronics')
data['Product Group'] = data['Product Group'].str.replace('Power', 'Power ')
data['Product Group'] = data['Product Group'].str.replace('Car', 'Car ')
data['Product Group'] = data['Product Group'].str.replace('Medical', ' Medical ')
data['Product Group'] = data['Product Group'].str.replace('Audio', ' Audio')
```

In [22]:
```
# View data grouped by Product Division and Product Group
grouping = data.groupby(['Product Division', 'Product Group'])
grouping.first()
```

Out[22]:

Product Division	Product Group	Product ID	Product Name	Category	Discounted Price	Actual Price	Discou Percentag
Car & Motorbike	**Car Accessories**	B0912WJ87V	Reffair AX30 [MAX] Portable Air Purifier for C...	Car&Motorbike|CarAccessories|InteriorAccessori...	2339.0	4000.0	42
Computers & Accessories	**Accessories & Peripherals**	B07JW9H4J1	Wayona Nylon Braided USB to Lightning Fast Cha...	Computers&Accessories|Accessories&Peripherals|...	399.0	1099.0	64
	Components	B08C4Z69LN	Crucial RAM 8GB DDR4 3200MHz CL22 (or 2933MHz ...	Computers&Accessories|Components|Memory	1792.0	3500.0	49
	External Devices & Data Storage	B005FYNT3G	SanDisk Cruzer Blade 32GB USB Flash Drive	Computers&Accessories|ExternalDevices&DataStor...	289.0	650.0	56
	Laptops	B0B2RBP83P	Lenovo IdeaPad 3 11th Gen Intel Core i3 15.6" ...	Computers&Accessories|Laptops|TraditionalLaptops	37247.0	59890.0	38
	Monitors	B08L879JSN	Acer EK220Q 21.5 Inch (54.61 cm) Full HD (1920...	Computers&Accessories|Monitors	6299.0	13750.0	54
	Networking Devices	B008IFXQFU	TP-Link USB WiFi Adapter for PC(TL-WN725N), N1...	Computers&Accessories|NetworkingDevices|Networ...	499.0	999.0	50
	Printers, Inks & Accessories	B08CYPB15D	HP 805 Black Original Ink Cartridge	Computers&Accessories|Printers,Inks&Accessorie...	717.0	761.0	6

Product Division	Product Group	Product ID	Product Name	Category	Discounted Price	Actual Price	Discou Percentag
	Tablets	B09XXZXQC1	Xiaomi Pad 5| Qualcomm Snapdragon 860| 120Hz R...	Computers&Accessories|Tablets	26999.0	37999.0	29
Electronics	**Accessories**	B0BDRVFDKP	SanDisk Ultra® microSDXC™ UHS-I Card, 64GB, 14...	Electronics|Accessories|MemoryCards|MicroSD	569.0	1000.0	43
	Cameras & Photography	B08FTFXNNB	HP w100 480P 30 FPS Digital Webcam with Built-...	Electronics|Cameras&Photography|VideoCameras	499.0	1999.0	75
	General Purpose Batteries & Battery Chargers	B014SZO90Y	Duracell Ultra Alkaline AA Battery, 8 Pcs	Electronics|GeneralPurposeBatteries&BatteryCha...	266.0	315.0	16
	Headphones, Earbuds & Accessories	B01DEWVZ2C	JBL C100SI Wired In Ear Headphones with Mic, J...	Electronics|Headphones,Earbuds&Accessories|Hea...	599.0	999.0	40
	Home Audio	B08L4SBJRY	Saifsmart Outlet Wall Mount Hanger Holder for ...	Electronics|HomeAudio|Accessories|SpeakerAcces...	349.0	1299.0	73
	Home Theater, TV & Video	B07KSMBL2H	AmazonBasics Flexible Premium HDMI Cable (Blac...	Electronics|HomeTheater,TV&Video|Accessories|C...	219.0	700.0	69
	Mobiles & Accessories	B08HV83HL3	MI Power Bank 3i 20000mAh Lithium Polymer 18W ...	Electronics|Mobiles&Accessories|MobileAccessor...	2049.0	2199.0	7

Product Division	Product Group	Product ID	Product Name	Category	Discounted Price	Actual Price	Discou Percentag
	Power Accessories	B0083T231O	Belkin Essential Series 4-Socket Surge Protect...	Electronics|PowerAccessories|SurgeProtectors	1289.0	1499.0	14
	Wearable Technology	B0BF57RN3K	Fire-Boltt Ninja Call Pro Plus 1.83" Smart Wat...	Electronics|WearableTechnology|SmartWatches	1799.0	19999.0	91
Health & Personal Care	**Home Medical Supplies & Equipment**	B07BKSSDR2	Dr Trust Electronic Kitchen Digital Scale Weig...	Health&PersonalCare|HomeMedicalSupplies&Equipm...	899.0	1900.0	53
Home & Kitchen	**Craft Materials**	B00N1U9AJS	3M Scotch Double Sided Heavy Duty Tape(1m hold...	Home&Kitchen|CraftMaterials|Scrapbooking|Tape	130.0	165.0	21
	Heating, Cooling & Air Quality	B00H47GVGY	USHA Quartz Room Heater with Overheating Prote...	Home&Kitchen|Heating,Cooling&AirQuality|RoomHe...	1199.0	1695.0	29
	Home Storage & Organization	B0814P4L98	PrettyKrafts Laundry Basket for clothes with L...	Home&Kitchen|HomeStorage&Organization|LaundryO...	351.0	999.0	65
	Kitchen & Dining	B01LWYDEQ7	Pigeon Polypropylene Mini Handy and Compact Ch...	Home&Kitchen|Kitchen&Dining|KitchenTools|Manua...	199.0	495.0	60
	Kitchen & Home Appliances	B07WMS7TWB	Pigeon by Stovekraft Amaze Plus Electric Kettl...	Home&Kitchen|Kitchen&HomeAppliances|SmallKitch...	649.0	1245.0	48

Product Division	Product Group	Product ID	Product Name	Category	Discounted Price	Actual Price	Discou Percentag
Home Improvement	**Electrical**	B07WKBD37W	ESnipe Mart Worldwide Travel Adapter with Buil...	HomeImprovement|Electrical|Adapters&Multi-Outlets	425.0	999.0	57
Musical Instruments	**Microphones**	B076B8G5D8	Boya ByM1 Auxiliary Omnidirectional Lavalier C...	MusicalInstruments|Microphones|Condenser	798.0	1995.0	60
Office Products	**Office Electronics**	B0846D5CBP	Casio FX-991ES Plus-2nd Edition Scientific Cal...	OfficeProducts|OfficeElectronics|Calculators|S...	1295.0	1295.0	0
	Office Paper Products	B07KCMR8D6	Classmate Octane Neon-Blue Gel Pens(Pack of 5...	OfficeProducts|OfficePaperProducts|Paper|Stati...	50.0	50.0	0
Toys & Games	**Arts & Crafts**	B00DJ5N9VK	Faber-Castell Connector Pen Set - Pack of 25 (...	Toys&Games|Arts&Crafts|Drawing&PaintingSupplie...	150.0	150.0	0

29 rows × 21 columns

In [23]:
```
# Extract required column for data visualization and further determination on relationship between variables
finaldata = data[['Product ID', 'Product Name', 'Product Division', 'Product Group', 'Actual Price', 'Discounted Price'
finaldata = finaldata.sort_values('Actual Price',ascending=False)
```

In [24]:
```
# Export excel file
finaldata.to_csv('amazonsales.csv',index=False)
```