GROUP MEMBERS ALONG WITH THERE ID'S

1. **MUHAMMAD NOMAN (SP22-MSCS-0013)**
2. **FIZA MEMON (SP22-MSCS-0018)**
3. **MEHRAB AZAM (SP22-MSCS-0047)**

SUBJECT: **DATA SCIENCE**

ASSIGNMENT TOPICS: Mid Project Report

DATE: **05/12/2023**

SUBMITTED TO **DR SYED IMRAN JAMI**

Project Idea:

E-commerce Data Analysis from Priceoye.pk

Objective: Our primary goal is to gain insights into product popularity and consumer behavior on Priceoye.pk, a prominent e-commerce website. Additionally, we plan to represent these insights using graphical visualizations.

BUSINESS QUESTION THAT WE ARE GOING TO SOLVE

The main business problems that our particular data science research seeks to answer have to do with figuring out how customers use the Priceoye.pk e-commerce platform and what products are popular. We aim to learn more about the products that are selling well, the reasons that drive customer choice, and the effects that price, discounts, brands, product categories, and product pictures have on consumer behavior.

Approach:

1. **Data Collection:** Develop web scraping scripts to gather comprehensive data on product listings, attributes, prices, user reviews, and ratings from Priceoye.pk.
2. **Data Analysis:** Employ data science methodologies to clean and analyze the collected data. This will include exploratory data analysis, statistical analysis, and machine learning techniques.
3. **Visualization:** Create graphical visualizations, such as charts and graphs, to represent the insights derived from the data. Visualization enhances understanding and communication.
4. **Project Presentation:** Prepare a project presentation to showcase our findings. This will include a summary of insights, visual representations, and explanations.

We don’t have the products count so that we can not judge which is more in quantity, currently we can’t check which item is selling more

Task Distribution

Assumptions:

Assumption of data is all upon the data which is scraped from website

Bug in Project:

There is no bug in this project

Code:

Libraries:

```
import requests
import matplotlib.pyplot as plt
import pandas as pd
```

```
from bs4 import BeautifulSoup
page = requests.get("https://priceoye.pk/")
page.status_code
```

Website Parsing:

```
soup = BeautifulSoup(page.content, 'html.parser')
```

Category Navigations:

```
navigation_div = soup.find_all('div', class_='sb-all-category')[0]
navigation_div
navigation_div.find_all('a')[0]['href']
```

Page Navigation Links:

```
navigation_links = navigation_div.find_all('a')
navigation = []
product_boxes = []

for i in navigation_links:
    #page_link = i['href']
    #sublink_page = requests.get(page_link)
    #sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
    #product_boxes.append(sublink_soup.find_all('div',class_='productBox')
)
    navigation.append(i['href'])
navigation
```

Sub Links:

```
page_link = navigation[0]
sublink_page = requests.get(page_link)
sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
sublink_soup
```

Product Count:

```
product_boxes = sublink_soup.find_all('div',class_='productBox')
len(product_boxes)
```

Product Name:

```
product_boxes[0].find('h4').text.strip()
```

Dataset into Dataframe:

```
product_name = product_boxes[0].find('h4').text.strip()
price = product_boxes[0].find('div',class_='price-box').text.strip().repla
ce('Rs. ','').replace(',','')
retail=product_boxes[0].find('div',class_='price-diff-retail').text.strip(
).replace('Rs. ','').replace(',','')
image_url = product_boxes[0].find('img')['src']
discount = product_boxes[0].find('div',class_='price-diff-saving').text.st
rip().replace('Rs. ','').replace(',','').replace(' OFF','')
brand = product_boxes[0]['data-brand']
category = product_boxes[0].find('a')['href'].split('https://priceoye.pk/'
)[1].split('/')[0]
```

```
product_boxes[0].find('h4').text.strip()
product_boxes[0].find('a')['href'].split('https://priceoye.pk/')[1]
```

```
import pandas as pd
df = pd.DataFrame(columns=['Product Name','Price','Retail Price','Discount
','Brand','Category','Image'])


for p in range(len(product_boxes)-1):
    try:
        product_name = product_boxes[p].find('h4').text.strip()
        price = product_boxes[p].find('div',class_='price-box').text.strip
().replace('Rs. ','').replace(',','')
        retail=product_boxes[p].find('div',class_='price-diff-retail').tex
t.strip().replace('Rs. ','').replace(',','')
        image_url = product_boxes[p].find('img')['src']
        brand = product_boxes[p]['data-brand']
        discount = product_boxes[p].find('div',class_='price-diff-saving')
.text.strip().replace('Rs. ','').replace(',','').replace(' OFF','')
        category = product_boxes[p].find('a')['href'].split('https://price
oye.pk/')[1].split('/')[0]
    except:
        pass
    df = df.append({'Product Name':product_name , 'Price':price, 'Retail P
rice':retail,'Discount':discount,'Brand':brand,'Category':category,'Image'
:image_url}, ignore_index = True)

df
```

```
pagination = sublink_soup.find('div',class_='pagination')
all_pages = pagination.find_all('a')
all_links = []
for i in range(2,int(all_pages[-2].text)+1,1):
    all_links.append("{}?page={}".format(page_link,i))

for i in all_links:
    sublink_page = requests.get(i)
    sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
    product_boxes = sublink_soup.find_all('div',class_='productBox')
    for p in range(len(product_boxes)-1):
        try:
            product_name = product_boxes[p].find('h4').text.strip()
```

```
            price = product_boxes[p].find('div',class_='price-box').text.s
trip().replace('Rs. ','').replace(',','')
            retail=product_boxes[p].find('div',class_='price-diff-retail')
.text.strip().replace('Rs. ','').replace(',','')
            image_url = product_boxes[p].find('img')['src']
            brand = product_boxes[p]['data-brand']
            discount = product_boxes[p].find('div',class_='price-diff-savi
ng').text.strip().replace('Rs. ','').replace(',','').replace(' OFF','')
            category = product_boxes[p].find('a')['href'].split('https://p
riceoye.pk/')[1].split('/')[0]
        except:
            pass
        df = df.append({'Product Name':product_name , 'Price':price, 'Reta
il Price':retail,'Discount':discount,'Brand':brand,'Category':category,'Im
age':image_url}, ignore_index = True)
df
```

```
df = pd.DataFrame(columns=['Product Name','Price','Retail Price','Discount
','Brand','Image'])
for link in navigation:
    sublink_page = requests.get(link)
    sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
    product_boxes = sublink_soup.find_all('div',class_='productBox')
    for p in range(len(product_boxes)-1):
        try:
            product_name = product_boxes[p].find('h4').text.strip()
            price = product_boxes[p].find('div',class_='price-box').text.s
trip().replace('Rs. ','').replace(',','')
            retail=product_boxes[p].find('div',class_='price-diff-retail')
.text.strip().replace('Rs. ','').replace(',','')
            image_url = product_boxes[p].find('img')['src']
            brand = product_boxes[p]['data-brand']
            discount = product_boxes[p].find('div',class_='price-diff-savi
ng').text.strip().replace('Rs. ','').replace(',','').replace(' OFF','')
            category = product_boxes[p].find('a')['href'].split('https://p
riceoye.pk/')[1].split('/')[0]
        except:
            pass
        df = df.append({'Product Name':product_name , 'Price':price, 'Reta
il Price':retail,'Discount':discount,'Brand':brand,'Category':category,'Im
age':image_url}, ignore_index = True)

print(df)
```

All categories Count:

```
df['Category'].value_counts()
```

Saving Data into Excel:

```
df.to_csv('products_first_page.csv')
```

Dataframe:

```
df = pd.DataFrame(columns=['Product Name','Price','Retail Price','Discount
','Brand','Image'])
for link in navigation:
    sublink_page = requests.get(link)
    sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
    product_boxes = sublink_soup.find_all('div',class_='productBox')
    for p in range(len(product_boxes)-1):
        try:
            product_name = product_boxes[p].find('h4').text.strip()
            price = product_boxes[p].find('div',class_='price-box').text.s
trip().replace('Rs. ','').replace(',','')
            retail=product_boxes[p].find('div',class_='price-diff-retail')
.text.strip().replace('Rs. ','').replace(',','')
            image_url = product_boxes[p].find('img')['src']
            brand = product_boxes[p]['data-brand']
            discount = product_boxes[p].find('div',class_='price-diff-savi
ng').text.strip().replace('Rs. ','').replace(',','').replace(' OFF','')
            category = product_boxes[p].find('a')['href'].split('https://p
riceoye.pk/')[1].split('/')[0]
        except:
            pass
        df = df.append({'Product Name':product_name , 'Price':price, 'Reta
il Price':retail,'Discount':discount,'Brand':brand,'Category':category,'Im
age':image_url}, ignore_index = True)
    #finding navigation on each link
    try:
      pagination = sublink_soup.find('div',class_='pagination')
      all_pages = pagination.find_all('a')
      all_links = []
    except:
      pass
    for i in range(2,int(all_pages[-2].text)+1,1):
        all_links.append("{}?page={}".format(link,i))
    for i in all_links:
        sublink_page = requests.get(i)
        sublink_soup = BeautifulSoup(sublink_page.content, 'html.parser')
        product_boxes = sublink_soup.find_all('div',class_='productBox')
```

```
        for p in range(len(product_boxes)-1):
            try:
                product_name = product_boxes[p].find('h4').text.strip()
                price = product_boxes[p].find('div',class_='price-box').te
xt.strip().replace('Rs. ','').replace(',','')
                retail=product_boxes[p].find('div',class_='price-diff-reta
il').text.strip().replace('Rs. ','').replace(',','')
                image_url = product_boxes[p].find('img')['src']
                brand = product_boxes[p]['data-brand']
                discount = product_boxes[p].find('div',class_='price-diff-
saving').text.strip().replace('Rs. ','').replace(',','').replace(' OFF','')
                category = product_boxes[p].find('a')['href'].split('https
://priceoye.pk/')[1].split('/')[0]
            except:
                pass
            df = df.append({'Product Name':product_name , 'Price':price, '
Retail Price':retail,'Discount':discount,'Brand':brand,'Category':category
,'Image':image_url}, ignore_index = True)

print(df)


df.to_csv('priceoye_products.csv')
```

Dropping Unnamed values:

```
df.drop('Unnamed: 0', inplace=True, axis=1)
```

Category Count:

```
df['Category'].value_counts()
```

Brand Count:

```
df['Brand'].value_counts()
```

Complete Info:

```
df.info()
```

Complete Description:

```
df.describe()
```

Visualization of Category:

```
plt.figure(figsize=(15,6))
df.Category.value_counts().plot(kind='barh')
```

Visualization of Brand:

```
plt.figure(figsize=(15,10))
df.Brand.value_counts().plot(kind='barh', color='green')
```

Dataset:

1. Product Name
2. Price
3. Retail Price
4. Discount
5. Brand
6. Image
7. Category

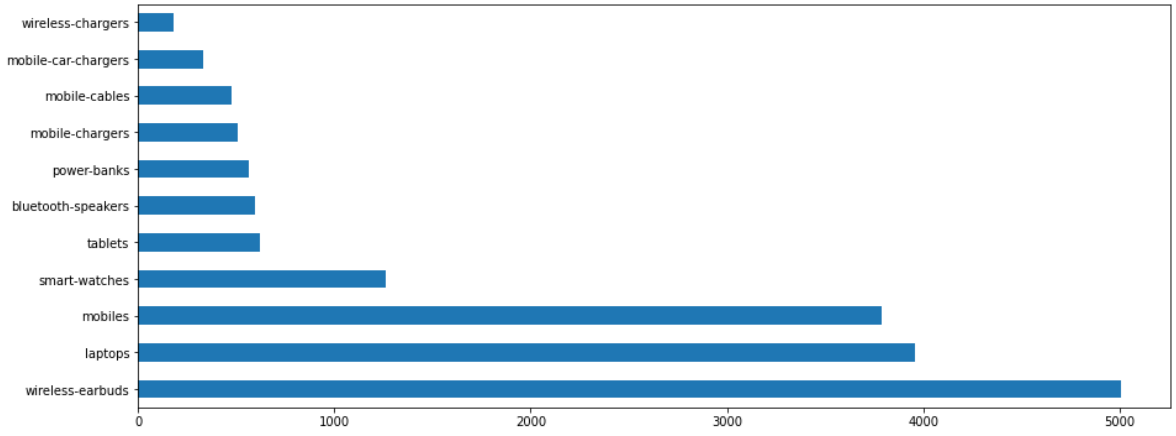
Summary of Data using Visualization:

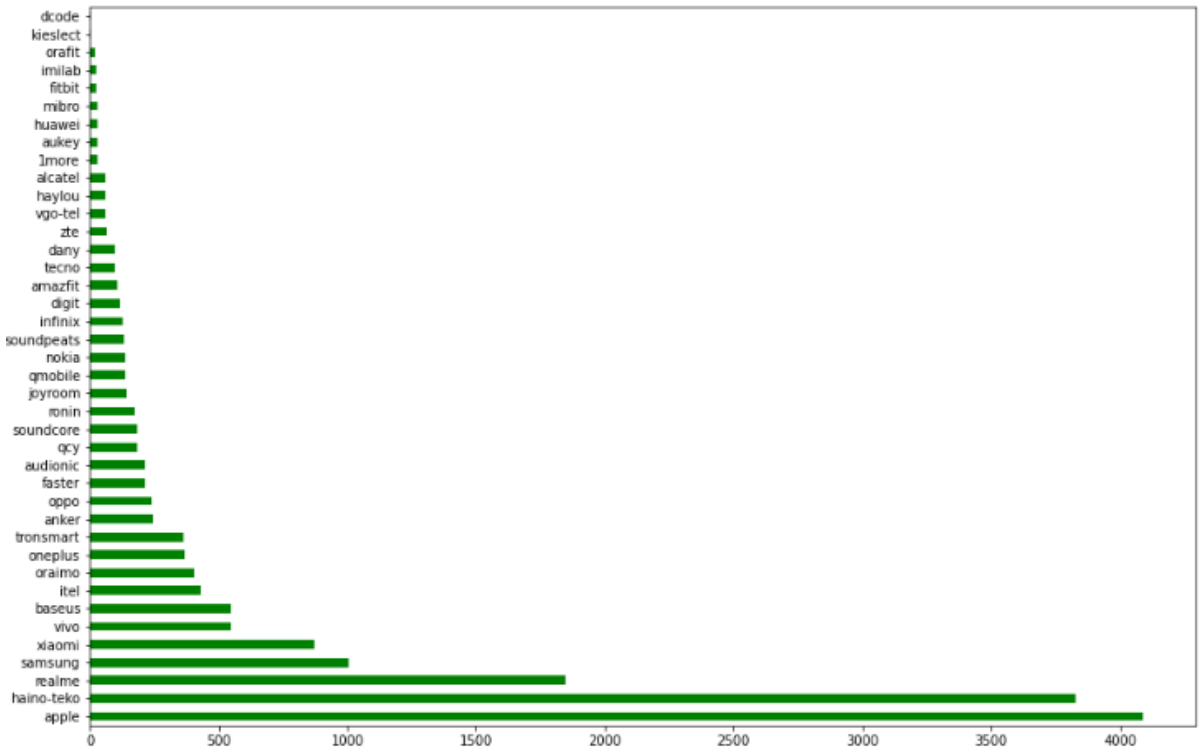
```
apple         4091
haino-teko    3826
realme        1848
samsung       1003
xiaomi         871
vivo           548
baseus         547
itel           429
oraimo         403
oneplus        368
tronsmart      365
anker          246
oppo           239
faster         217
audionic       215
qcy            185
soundcore      184
ronin          175
joyroom        142
qmobile        140
nokia          139
soundpeats     130
infinix        128
digit          115
amazfit        108
tecno           99
dany            94
zte             64
vgo-tel         62
haylou          61
alcatel         59
1more           32
aukey           32
huawei          32
mibro           28
fitbit          23
imilab          23
orafit          21
kieslect         5
dcode            4


laptops               1434
mobiles                862
wireless-earbuds       210
mobile-cables          139
smart-watches          106
bluetooth-speakers      98
mobile-chargers         93
power-banks             92
mobile-car-chargers     83
tablets                 50
wireless-chargers       41
Name: Category, dtype: int64
```

	Product Name	Price	Retail Price	Discount	Brand	Image	Category
0	Samsung Galaxy A32	38199	43999	5800	samsung	https://static.priceoye.pk/images/placeholder-...	mobiles
1	Infinix Hot 11 Play	20099	21999	1900	infinix	https://static.priceoye.pk/images/placeholder-...	mobiles
2	Samsung Galaxy A12	23199	24999	1800	samsung	https://static.priceoye.pk/images/placeholder-...	mobiles
3	Infinix Note 10 Pro	32399	34999	2600	infinix	https://static.priceoye.pk/images/placeholder-...	mobiles
4	Xiaomi Redmi Note 11	32999	34999	2000	xiaomi	https://static.priceoye.pk/images/placeholder-...	mobiles
...	...	...	...	...	...	...	...
3203	United US100 Jazba	59000	220000	5500	apple	https://static.priceoye.pk/images/product-plac...	laptops
3204	United US125 Deluxe	137500	220000	5500	apple	https://static.priceoye.pk/images/product-plac...	laptops
3205	United US150	142000	220000	5500	apple	https://static.priceoye.pk/images/product-plac...	laptops
3206	United US100	74500	220000	5500	apple	https://static.priceoye.pk/images/product-plac...	laptops
3207	ZXMCO ZX70	40500	220000	5500	apple	https://static.priceoye.pk/images/product-plac...	laptops

3208 rows × 7 columns





Project Learning Experience:

By this project we have learned many things. Our programming skills enhanced. Vision of python development increases. Apply data science skills to address a real-world problem .Visualization of the data is much useful for us in different things.