

Q. Use scrapy to crawl any of the E-commerce website of your choice.

The following information needs to be extracted from the page for the corresponding product:

- a) Product Name
- b) Product Price
- c) Product Discount
- d) Product Image (URL)

PROCEDURE:

A. SETUP

1. Installing scrapy library onto the device or virtual environment.

Enter in the terminal:

```
pip install scrapy
```

2. Move to the folder where you want to create your scrapy project in the terminal and enter:

```
scrapy startproject ecomscrape
```

to create a scrapy project with the name ecomscrape at the location

3. Move up to the root directory of the scrapy project using terminal command:

```
cd ecomscrape
```

4. In the ecomscrape directory of the root scrapy directory, another directory named spiders will be present where we have to create the spider we want to use.
Create a python3 file in the spiders directory with a name of your choice.
(ecomSpider.py here)

B. PROGRAMMING

5. Open the ecomSpider.py file in an IDE.
6. First we want to import the scrapy library using:

```
import scrapy
```

7. Create a class using the scrapy library with scrapy.Spider class

```
class PostsSpider(scrapy.Spider):
```

8. Inside the class, assign the spider we are creating an unique name as an identifier to call the spider later using terminal as:

```
name="posts"
```

9. Now, specify the list of base urls from where you want the scraper to start crawling as:

```
start_urls=["https://www.flipkart.com/search?q=earphones"]
```

10. Now, create a 'parse' function inside the class. The parse function is the default function which is called when the spider is run.

11. Type your program for data extraction inside the parse function. (use scrapy shell 'URL' command in terminal and test your snippets of logic in there to check if any particular query works or not)
12. Run the program to compile it for any errors. (This is not execution of the spider, its just compilation of the program to save it and check it for errors)

C. EXECUTION

13. If the robots.txt file of the website disallow the domain you want to access then in the settings.py file of the scrapy project set ROBOTSTXT_OBEY Boolean value to false and proceed. (It might be against the sites policy so check for it beforehand)
14. In the terminal enter the command:

```
scrapy crawl posts -o filename.json
```

15. The json file will be created with the required data in the root directory of the scrapy project.

CODE:

```
import scrapy
```

```
class PostsSpider(scrapy.Spider):
```

```
    name= "posts"
```

```
    start_urls=["https://www.flipkart.com/search?q=earphones"]
```

```
    def parse(self, response):
```

```
        for p in response.css('div._4ddWXP'):
```

```
            yield{
```

```
                'Name': p.css('a.s1Q9rs::text').get(),
```

```
                'Price': p.css('a._8VNY32 div._25b18c div._30jeq3::text').get().split("\u20b9")[1],
```

```
                'Images': p.css('img::attr(src)')[0].get(),
```

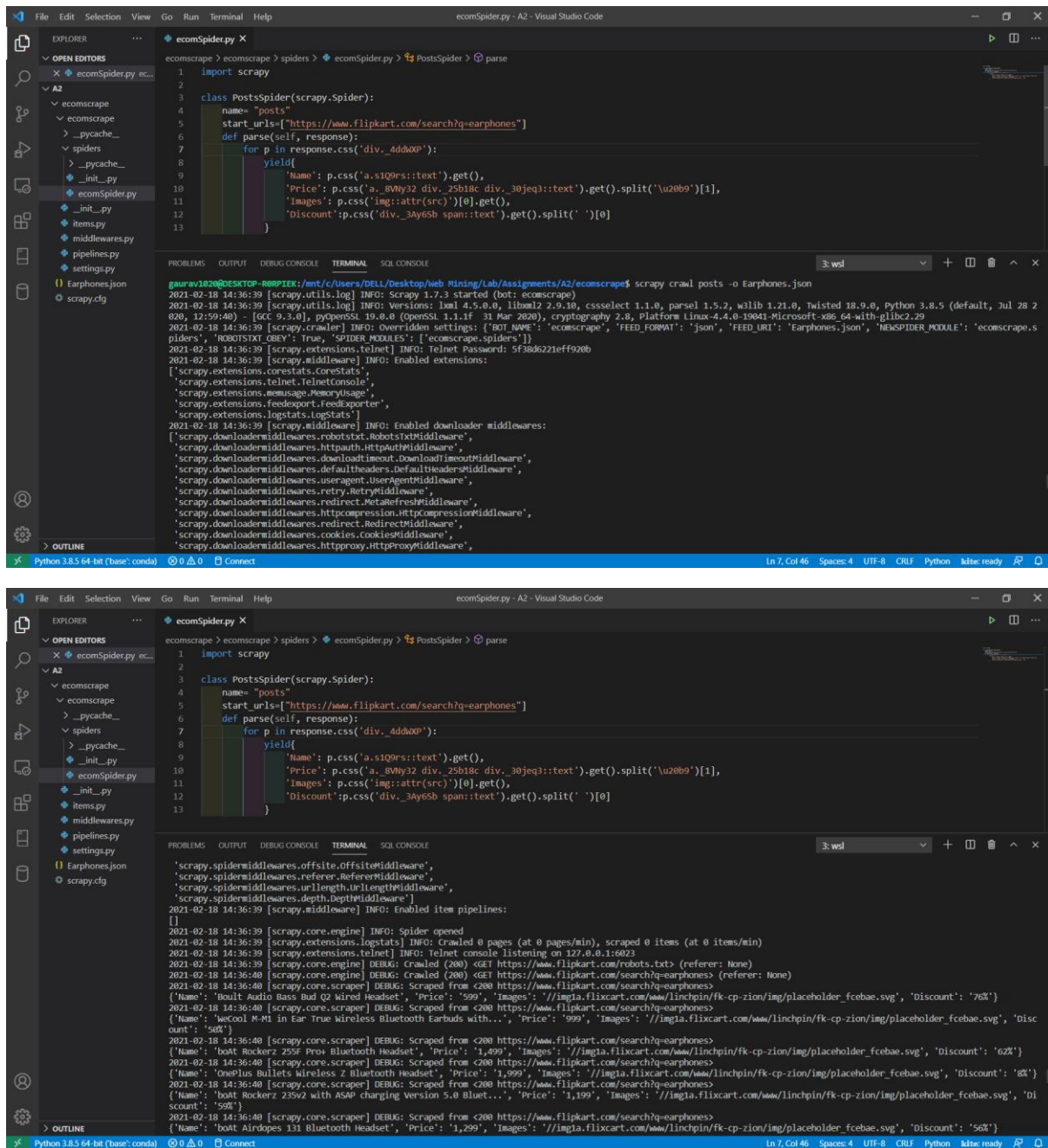
```
                'Discount':p.css('div._3Ay6Sb span::text').get().split(' ')[0]
```

```
            }
```

Registration Number:- 19BCE2119
Course:- Web Mining (L5+L6)

Name:- Gaurav Kumar Singh

OUTPUT:



```
ecomSpider.py X
ecomscrape > ecomscrape > spiders > ecomSpider.py > PostsSpider > parse
1 import scrapy
2
3 class PostsSpider(scrapy.Spider):
4     name = "posts"
5     start_urls = ["https://www.flipkart.com/search?q=earphones"]
6     def parse(self, response):
7         for p in response.css('div._4ddh0P'):
8             yield{
9                 'Name': p.css('a.s1Q9rs::text').get(),
10                'Price': p.css('a._8VWY32 div._25b18c div._30jeq3::text').get().split('\u20B9')[1],
11                'Images': p.css('img::attr(src)')[0].get(),
12                'Discount': p.css('div._3Ay6Sb span::text').get().split(' ')[0]
13            }

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE
3: wsl

gaurav102@DESKTOP-ND0PTEK: /mnt/c/Users/DELL/Desktop/web_Mining/Lab/Assignments/A2/ecomscrape$ scrapy crawl posts -o Earphones.json
2021-02-18 14:36:39 [scrapy.utils.log] INFO: Scrapy 1.7.3 started (bot: ecomscrape)
2021-02-18 14:36:39 [scrapy.utils.log] INFO: Versions: lxml 4.5.0.0, libxml2 2.9.10, cssselect 1.1.0, parsel 1.5.2, w3lib 1.21.0, Twisted 18.9.0, Python 3.8.5 (default, Jul 28 2
020, 12:59:40) - [GCC 9.3.0], pyOpenSSL 19.0.0 (OpenSSL 1.1.1f 31 Mar 2020), cryptography 2.8, Platform Linux-4.4.0-19041-Microsoft-x86_64-with-glibc2.29
2021-02-18 14:36:39 [scrapy.extensions.telnet] INFO: Telnet Password: 5f3bd6221eff920b
2021-02-18 14:36:39 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.message.retry.RetryUsage',
'scrapy.extensions.feedexport.FeedExporter',
'scrapy.extensions.logstats.LogStats']
2021-02-18 14:36:39 [scrapy.middleware] INFO: Enabled downloader middlewares:
['scrapy.downloadermiddlewares.robotstxt.RobotsTxtMiddleware',
'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware',
'scrapy.downloadermiddlewares.retry.RetryMiddleware',
'scrapy.downloadermiddlewares.redirect.MetaRefreshMiddleware',
'scrapy.downloadermiddlewares.httpcompression.HttpCompressionMiddleware',
'scrapy.downloadermiddlewares.redirect.RedirectMiddleware',
'scrapy.downloadermiddlewares.cookies.CookiesMiddleware',
'scrapy.downloadermiddlewares.httpproxy.HttpProxyMiddleware']

Python 3.8.5 64-bit (base: conda) 0 0 0 Connect Ln 7, Col 46 Spaces: 4 UTF-8 CRLF Python Idle ready
```

```
ecomSpider.py X
ecomscrape > ecomscrape > spiders > ecomSpider.py > PostsSpider > parse
1 import scrapy
2
3 class PostsSpider(scrapy.Spider):
4     name = "posts"
5     start_urls = ["https://www.flipkart.com/search?q=earphones"]
6     def parse(self, response):
7         for p in response.css('div._4ddh0P'):
8             yield{
9                 'Name': p.css('a.s1Q9rs::text').get(),
10                'Price': p.css('a._8VWY32 div._25b18c div._30jeq3::text').get().split('\u20B9')[1],
11                'Images': p.css('img::attr(src)')[0].get(),
12                'Discount': p.css('div._3Ay6Sb span::text').get().split(' ')[0]
13            }

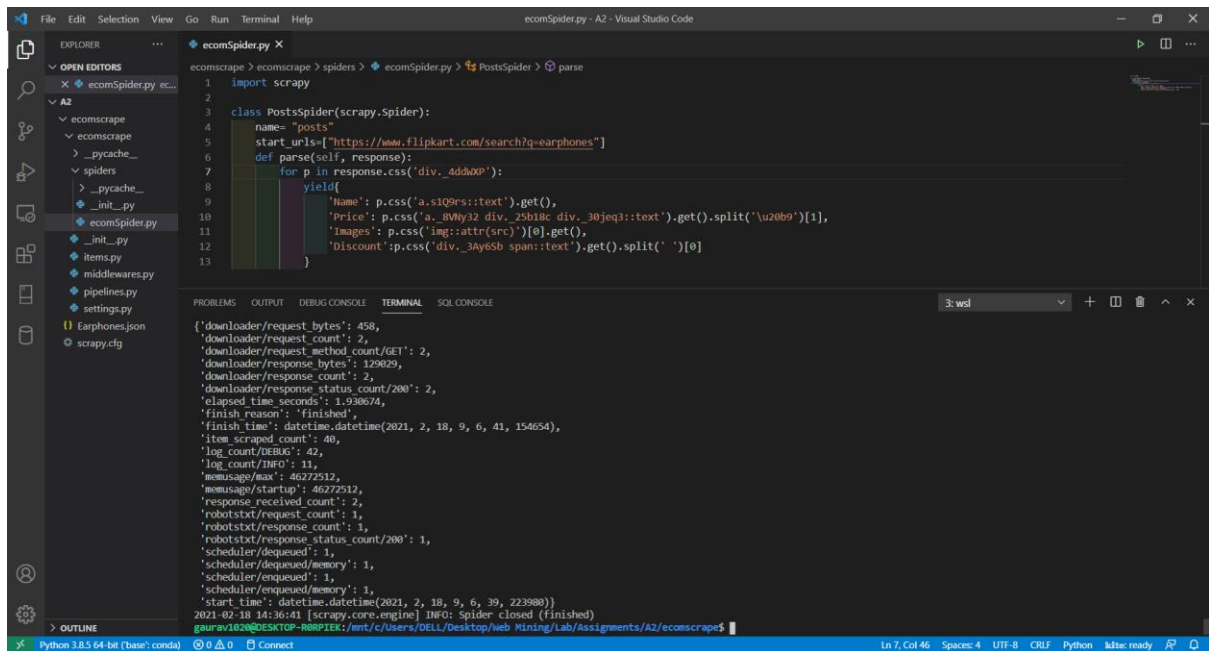
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL SQL CONSOLE
3: wsl

'scrapy.spidermiddlewares.offsite.OffsiteMiddleware',
'scrapy.spidermiddlewares.referer.RefererMiddleware',
'scrapy.spidermiddlewares.urllength.UrlLengthMiddleware',
'scrapy.spidermiddlewares.depth.DepthMiddleware']
2021-02-18 14:36:39 [scrapy.middleware] INFO: Enabled item pipelines:
[]
2021-02-18 14:36:39 [scrapy.core.engine] INFO: Spider opened
2021-02-18 14:36:39 [scrapy.extensions.logstats] INFO: Crawled 0 pages (at 0 pages/min), scraped 0 items (at 0 items/min)
2021-02-18 14:36:39 [scrapy.extensions.telnet] INFO: Telnet console listening on 127.0.0.1:6023
2021-02-18 14:36:39 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.flipkart.com/robots.txt> (referer: None)
2021-02-18 14:36:40 [scrapy.core.engine] DEBUG: Crawled (200) <GET https://www.flipkart.com/search?q=earphones> (referer: None)
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'Boult Audio Bass Bud Q2 Wired Headset', 'Price': '599', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Discount': '76%'}
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'Wecool M-M1 in Ear True Wireless Bluetooth Earbuds with...', 'Price': '599', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Disc
ount': '56%'}
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'boAt Rockerz 255F Prox Bluetooth Headset', 'Price': '1,499', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Discount': '62%'}
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'OnePlus Bullets Wireless Z Bluetooth Headset', 'Price': '1,999', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Discount': '6%'}
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'boAt Rockerz 235v2 with ASAP charging Version 5.0 Bluet...', 'Price': '1,199', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Di
scount': '59%'}
2021-02-18 14:36:40 [scrapy.core.scraper] DEBUG: Scraped from <200 https://www.flipkart.com/search?q=earphones>
{'Name': 'boAt Airdopes 131 Bluetooth Headset', 'Price': '1,299', 'Images': '//imgia.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg', 'Discount': '56%'}

Python 3.8.5 64-bit (base: conda) 0 0 0 0 Connect Ln 7, Col 46 Spaces: 4 UTF-8 CRLF Python Idle ready
```

Registration Number:- 19BCE2119
Course:- Web Mining (L5+L6)

Name:- Gaurav Kumar Singh

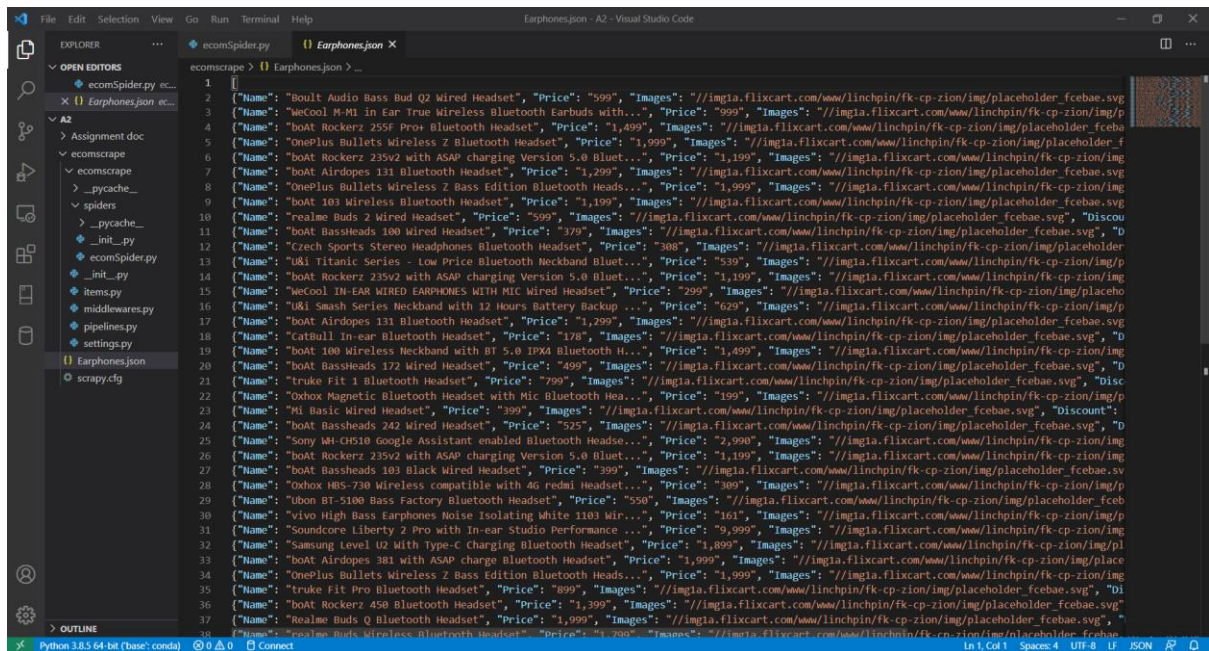


The screenshot shows a Visual Studio Code editor with the file `ecomSpider.py` open. The code defines a `PostsSpider` class that inherits from `scrapy.Spider`. It sets the name to 'posts' and the start_urls to `['https://www.flipkart.com/search?q=earphones']`. The `parse` method extracts product details from the response, including name, price, images, and discount. The terminal output shows the spider's execution details, including the start time, the number of items scraped (40), and the elapsed time (1.930674 seconds).

```
import scrapy

class PostsSpider(scrapy.Spider):
    name = "posts"
    start_urls = ['https://www.flipkart.com/search?q=earphones']
    def parse(self, response):
        for p in response.css('div._4dd00p'):
            yield {
                'name': p.css('a.s10Qrss:text').get(),
                'price': p.css('a._8uW32 div._250aie div._30jeq3:text').get().split('\u20b9')[1],
                'images': p.css('img::attr(src)')[0].get(),
                'discount': p.css('div._3Ay6Sb span::text').get().split(' ')[0]
```

```
{'download/request_bytes': 458,
'download/request_count': 2,
'download/request_method_count/GET': 2,
'download/response_bytes': 129029,
'download/response_count': 2,
'download/response_status_count/200': 2,
'elapsed_time_seconds': 1.930674,
'finish_reason': 'finished',
'finish_time': datetime.datetime(2021, 2, 18, 9, 6, 41, 154654),
'item_scraped_count': 40,
'log_count/DEBUG': 42,
'log_count/INFO': 11,
'memusage/max': 46272512,
'memusage/startup': 46272512,
'response_received_count': 2,
'robotstxt/request_count': 1,
'robotstxt/response_count': 1,
'robotstxt/response_status_count/200': 1,
'scheduler/dequeued': 1,
'scheduler/dequeued/memory': 1,
'scheduler/enqueued': 1,
'scheduler/enqueued/memory': 1,
'start_time': datetime.datetime(2021, 2, 18, 9, 6, 39, 223880)}
2021-02-18 14:36:41 [scrapy.core.engine] INFO: Spider closed (finished)
gaurav182@DESKTOP-R00PIEK: /mnt/c/Users/DELL/Desktop/web Mining/Lab/Assignments/A2/ecomscape$
```



The screenshot shows a Visual Studio Code editor with the file `Earphones.json` open. The file contains a list of 40 earphone products, each with its name, price, and image URL. The products are listed in a JSON array format.

```
[{"Name": "Boult Audio Bass Bud Q2 Wired Headset", "Price": "599", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "MeCool M-MI in Ear True Wireless Bluetooth Earbuds with...", "Price": "999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "boat Rockerz 255F Pro+ Bluetooth Headset", "Price": "1,499", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae"}, {"Name": "OnePlus Bullets Wireless Z Bluetooth Headset", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_f"}, {"Name": "boat Rockerz 235v2 with ASAP charging Version 5.0 Bluet...", "Price": "1,199", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "boat Airdopes 131 Bluetooth Headset", "Price": "1,299", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "OnePlus Bullets Wireless Z Bass Edition Bluetooth Heads...", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "boat 103 Wireless Bluetooth Headset", "Price": "1,199", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "realme Buds 2 Wired Headset", "Price": "599", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "boat BassHeads 100 Wired Headset", "Price": "379", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "Czech Sports Stereo Headphones Bluetooth Headset", "Price": "388", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder"}, {"Name": "U&I Titanic Series - Low Price Bluetooth Neckband Bluet...", "Price": "539", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "boat Rockerz 235v2 with ASAP charging Version 5.0 Bluet...", "Price": "1,199", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "McCool IN-EAR WIRED EARPHONES WITH MIC Wired Headset", "Price": "299", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placaho"}, {"Name": "U&I Smash Series Neckband with 12 Hours Battery Backup ...", "Price": "629", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "boat Airdopes 131 Bluetooth Headset", "Price": "1,299", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "CatBull In-ear Bluetooth Headset", "Price": "178", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "boat 100 Wireless Neckband with BT 5.0 IPX4 Bluetooth H...", "Price": "1,499", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "boat BassHeads 172 Wired Headset", "Price": "499", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "Truke Fit 1 Bluetooth Headset", "Price": "799", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "Oxbox Magnetic Bluetooth Headset with Mic Bluetooth Hea...", "Price": "199", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "MI Basic Wired Headset", "Price": "399", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "boat BassHeads 242 Wired Headset", "Price": "525", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "Sony WH-CH510 Google Assistant enabled Bluetooth Headse...", "Price": "2,990", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "boat Rockerz 235v2 with ASAP charging Version 5.0 Bluet...", "Price": "1,199", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "boat BassHeads 103 Black Wired Headset", "Price": "399", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.sv"}, {"Name": "Oxbox HBS-730 Wireless compatible with 4G redmi Headset...", "Price": "309", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "Ubton BT-5100 Bass Factory Bluetooth Headset", "Price": "550", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae"}, {"Name": "vivo High Bass Earphones Noise Isolating White 1103 Wir...", "Price": "161", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/p"}, {"Name": "Soundcore Liberty 2 Pro with In-ear Studio Performance ...", "Price": "9,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "Samsung Level U2 With Type-C charging Bluetooth Headset", "Price": "1,899", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/pl"}, {"Name": "boat Airdopes 381 with ASAP charge Bluetooth Headset", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/plac"}, {"Name": "OnePlus Bullets Wireless Z Bass Edition Bluetooth Heads...", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img"}, {"Name": "Truke Fit Pro Bluetooth Headset", "Price": "899", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "boat Rockerz 450 Bluetooth Headset", "Price": "1,399", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "realme buds Q Bluetooth Headset", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae.svg"}, {"Name": "OnePlus Bullets Wireless Z Bass Edition Bluetooth Headset", "Price": "1,999", "Images": "https://img1a.flixcart.com/www/linchpin/fk-cp-zion/img/placeholder_fcbae"}]
```