Kendriya Vidyalaya Malleshwaram

Project Title:

Web Scraping

Class 12 - B

Group - 18

Name	Roll No.	CBSE Roll No.
Hemish S	**12223**	
Pranav A S	**12233**	

CERTIFICATE

This is to certify that **Pranav A S** and **Hemish S** of class XII B have successfully completed the Computer Science project.

The project titled **“Web Scraping”** is the bonafide work of the above students as per the CBSE curriculum 2022-23

Internal Examiner

[Bhavana]
PGT - Computer Science

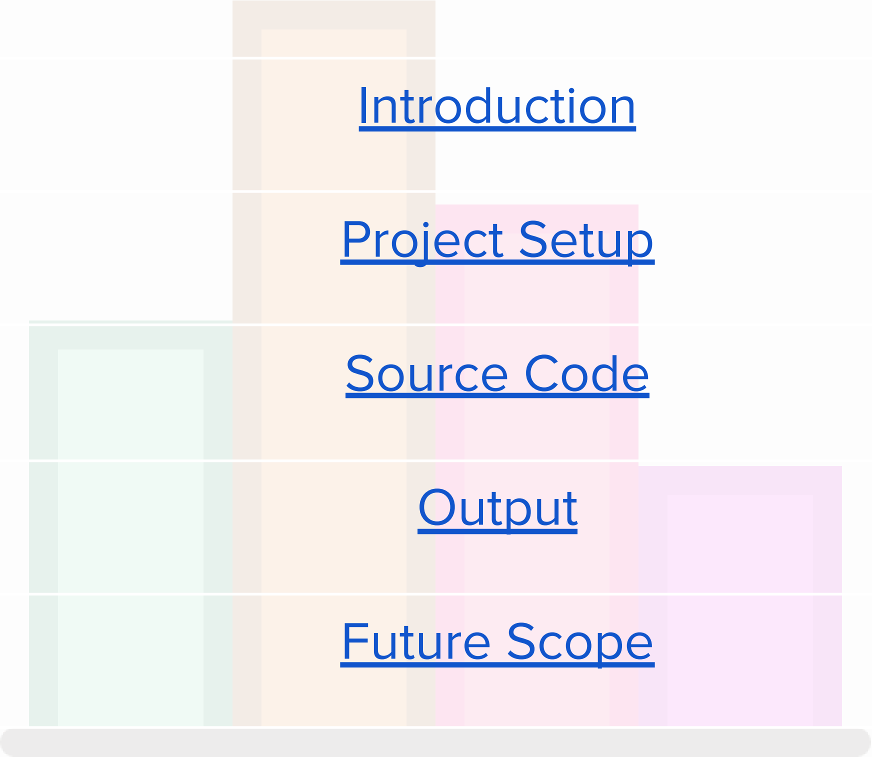
ACKNOWLEDGEMENT

We would like to convey our gratitude to our Computer Science teacher Ms. Bhavana for giving us the opportunity to do this project and for all her encouragement and guidance which lead to the successful completion of this project.

We would also like to thank our friends who helped us come up with the project idea and supported us throughout the development.

Finally we would like to thank each other for the cooperation and understanding.

TABLE OF CONTENTS



1.INTRODUCTION

What is Web Scraping?

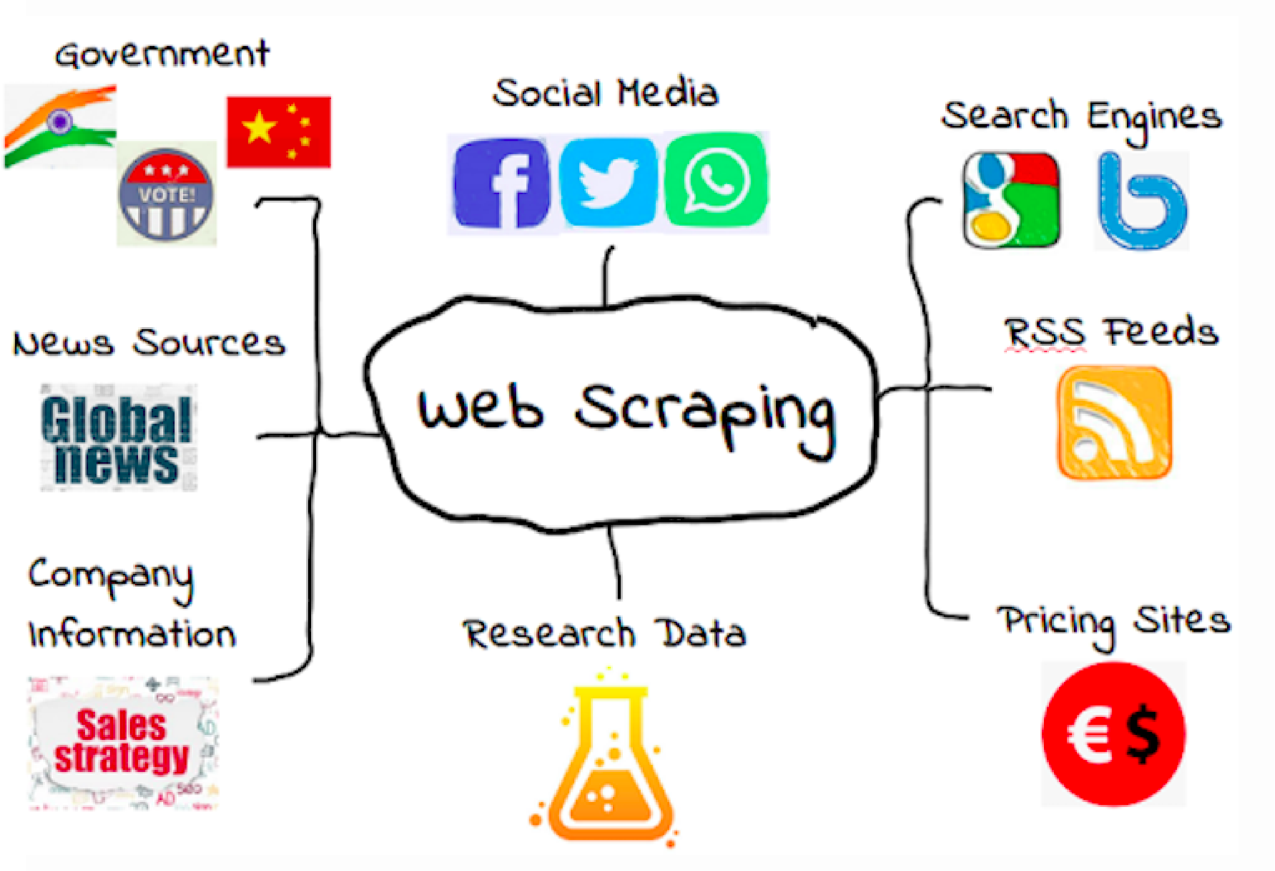
Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform web scraping to obtain data from websites.

How Web Scrapers Work?

Web Scrapers can extract all the data on particular sites or the specific data that a user wants. Ideally, it's best if you specify the data you want so that the web scraper only extracts that data quickly. For example, you might want to scrape an Amazon page for the types of juicers available, but you might only want the data about the models of different juicers and not the customer reviews. So, when a web scraper needs to scrape a site, first the URLs are provided. Then it loads all the HTML code for those sites and a more advanced scraper might even extract all the CSS and Javascript elements as well. Then the scraper obtains the required data from this HTML code and outputs this data in the format specified by the user. Mostly, this is in the form of an Excel spreadsheet or a CSV file, but the data can also be saved in other formats, such as a JSON file.

Why is Python a Popular Language for Web Scraping?

Python seems to be in fashion these days! It is the most popular language for web scraping as it can handle most of the processes easily. It also has a variety of libraries that were created specifically for Web Scraping. **Scrapy** is a very popular open-source web crawling framework that is written in Python. It is ideal for web scraping as well as extracting data using APIs. **Beautiful soup** is another Python library that is highly suitable for Web Scraping. It creates a parse tree that can be used to extract data from HTML on a website. Beautiful soup also has multiple features for navigation, searching, and modifying these parse trees. **We are going to be using Beautiful Soup for our Project.**



How are we going to use this?

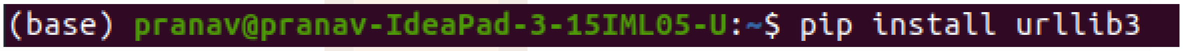
Through our project we want to show how web scraping works and how it can be used.
In this project we have built a Dataset collecting web scraper which collects all the search results from an **E-Commerce** website in our case let’s take **“Graphics Card”** with - **Name of the Product, Price, Rating, Stars & Image URL** and organizes the collected data into an **Excel sheet**.

This collected dataset can be used to further analyze the demand in the market for the given product or just even to find the best price for it.

2.PROJECT SETUP

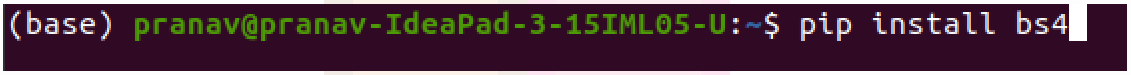
For coding this project we need to install the following modules and external libraries:

1. **urllib.request:**



```
(base) pranav@pranav-IdeaPad-3-15IML05-U:~$ pip install urllib3
```

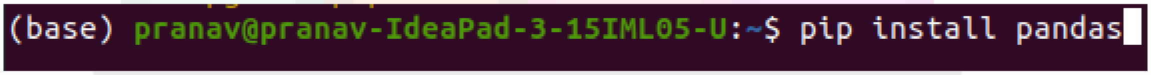
2. **bs4:**



```
(base) pranav@pranav-IdeaPad-3-15IML05-U:~$ pip install bs4
```

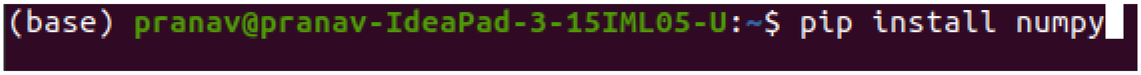
Optional:

3. **pandas:**



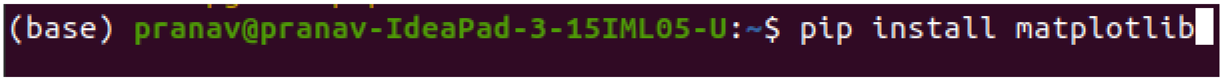
```
(base) pranav@pranav-IdeaPad-3-15IML05-U:~$ pip install pandas
```

4. **numpy:**



```
(base) pranav@pranav-IdeaPad-3-15IML05-U:~$ pip install numpy
```

5. **matplotlib:**



```
(base) pranav@pranav-IdeaPad-3-15IML05-U:~$ pip install matplotlib
```

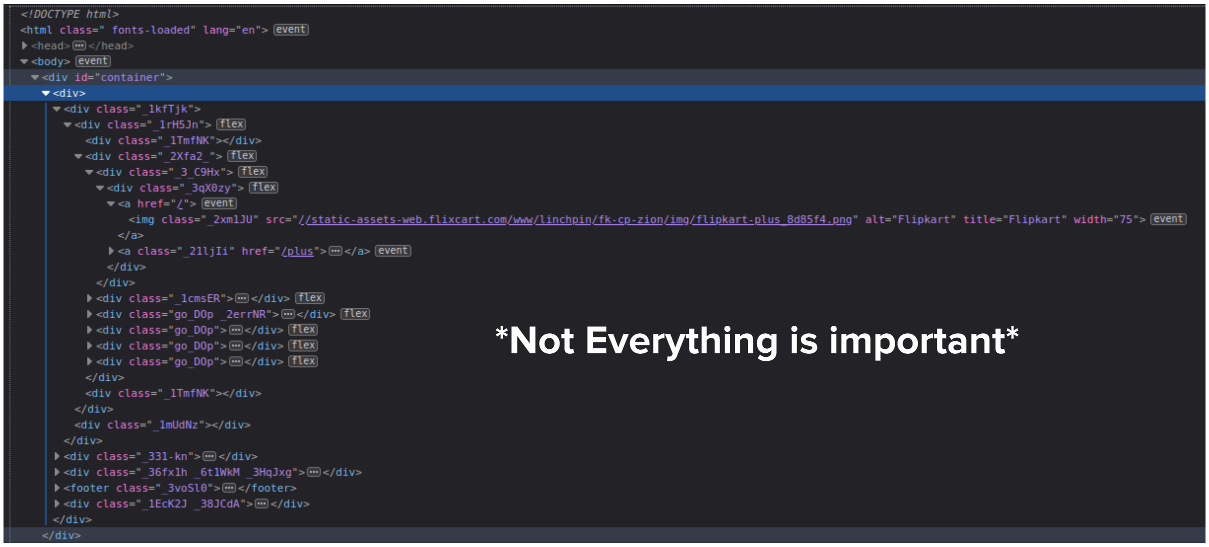
3.SOURCE CODE

IMPORTING THE NECESSARY LIBRARIES:

```
#importing libraries
from urllib.request import urlopen
from bs4 import BeautifulSoup as soup
import re
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
```

COLLECTING THE HTML DATA FROM THE WEBSITE

```
#reading the url
httpObject =
urlopen("https://www.flipkart.com/search?q=graphics+card")
webdata = httpObject.read()

soup1 = soup(webdata)
```

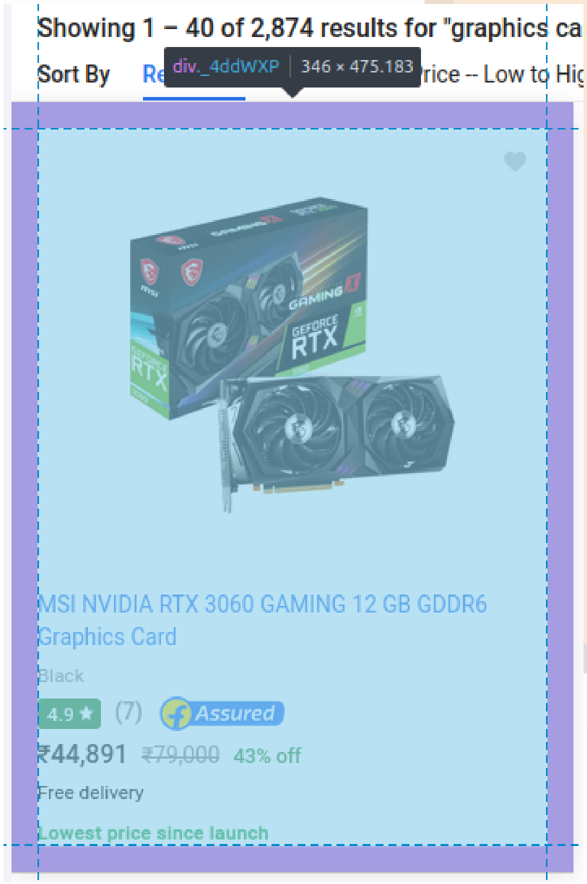


```
<!DOCTYPE html>
<html class=" fonts-loaded" lang="en"> event
▸ <head>…</head>
▾ <body> event
  ▾ <div id="container">
    ▾ <div>
      ▾ <div class="_1kfTjk">
        ▾ <div class="_1rH5Jn"> flex
            <div class="_1TmfNK"></div>
          ▾ <div class="_2Xfa2_"> flex
            ▾ <div class="_3_C9Hx"> flex
              ▾ <div class="_3qX0zy"> flex
                ▾ <a href="/"> event
                    <img class="_2xmlJU" src="//static-assets-web.flixcart.com/www/linchpin/fk-cp-zion/img/flipkart-plus_8d85f4.png" alt="Flipkart" title="Flipkart" width="75"> event
                  </a>
                ▸ <a class="_21ljIi" href="/plus">…</a> event
                </div>
              </div>
            ▸ <div class="_1cmsER">…</div> flex
            ▸ <div class="go_DOp _2errNR">…</div> flex
            ▸ <div class="go_DOp">…</div> flex
            ▸ <div class="go_DOp">…</div> flex
            ▸ <div class="go_DOp">…</div> flex
            </div>
            <div class="_1TmfNK"></div>
          </div>
          <div class="_1mUdNz"></div>
        </div>
      ▸ <div class="_331-kn">…</div>
      ▸ <div class="_36fxlh _6tlWkM _3HqJxg">…</div>
      ▸ <footer class="_3voSl0">…</footer>
      ▸ <div class="_1EcK2J _38JCdA">…</div>
      </div>
    </div>
```

Not Everything is important

FILTERING OUT THE UNWANTED HTML

```
pages_link = soup1.findAll('a',{'class':'ge-49M'})
domain = 'https://www.flipkart.com/search?q=graphics+card&page='
for i in range(2,11):
     link = domain+str(i)
     page_data = urlopen(link)
     webdata1 = page_data.read()
     webdata += webdata1

soupdata = soup(webdata, 'html.parser')
```



Showing 1 – 40 of 2,874 results for "graphics ca

MSI NVIDIA RTX 3060 GAMING 12 GB GDDR6 Graphics Card

Black

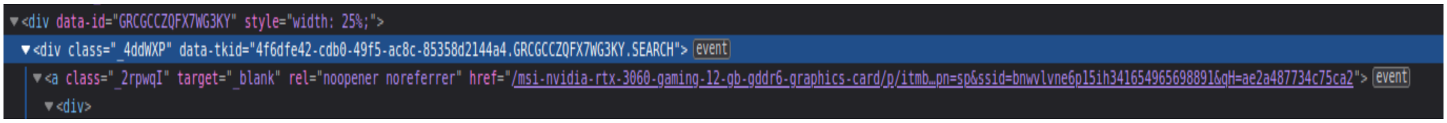
4.9 ★ (7) Assured

₹44,891 ~~₹79,000~~ 43% off

Free delivery

Lowest price since launch

IMPORTANT HTML



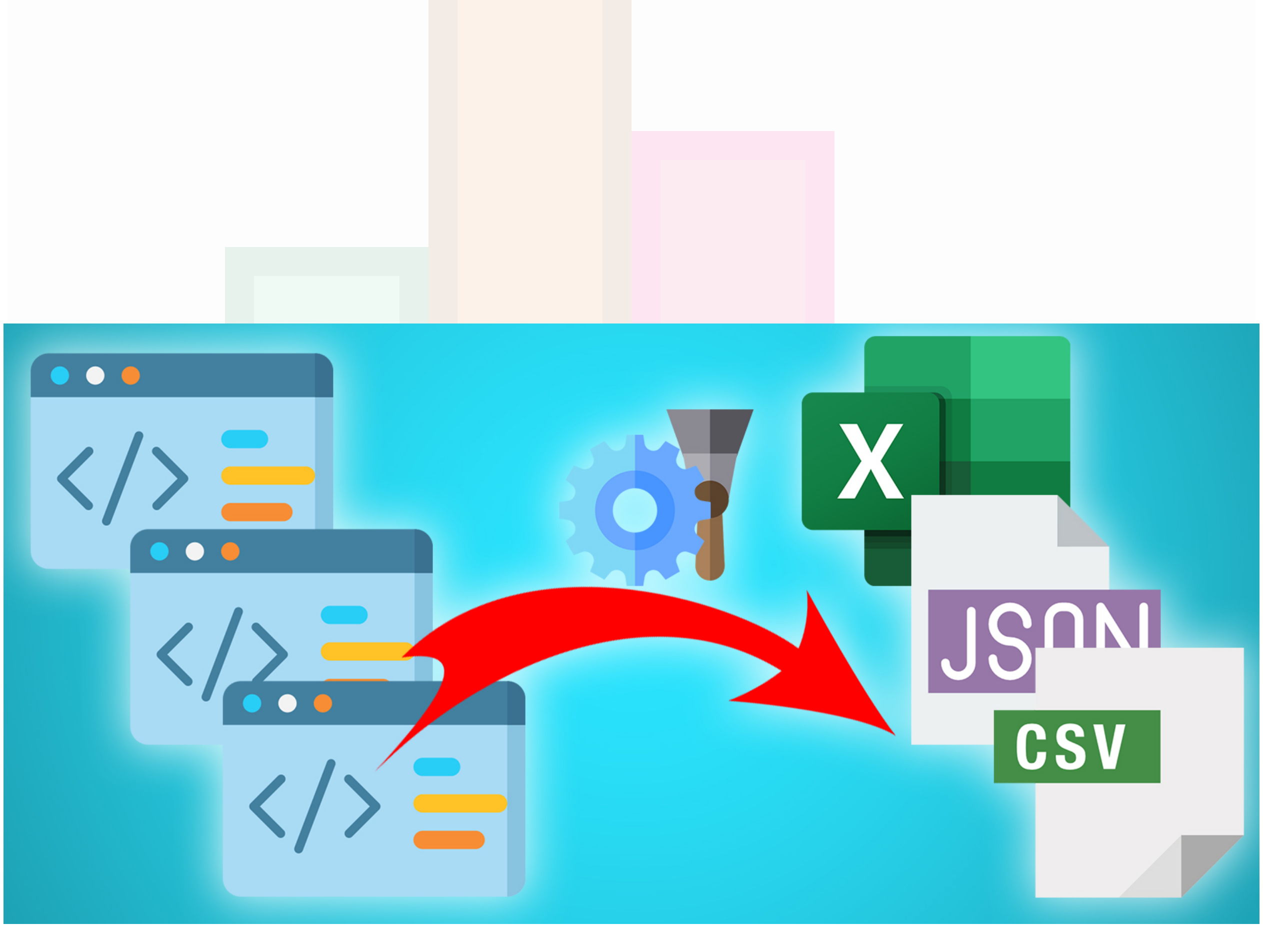
```
▼<div data-id="GRCGCCZQFX7WG3KY" style="width: 25%;">
 ▼<div class="_4ddWXP" data-tkid="4f6dfe42-cdb0-49f5-ac8c-85358d2144a4.GRCGCCZQFX7WG3KY.SEARCH"> event
  ▼<a class="_2rpwqI" target="_blank" rel="noopener noreferrer" href="/msi-nvidia-rtx-3060-gaming-12-gb-gddr6-graphics-card/p/itmb…pn=sp&ssid=bnwvlvne6p15ih341654965698891&qH=ae2a487734c75ca2"> event
   ▼<div>
```

ORGANIZING THE COLLECTED HTML

```
containers = soupdata.findAll('div',{'class':'_4ddWXP'})

print(type(containers),len(containers))
```

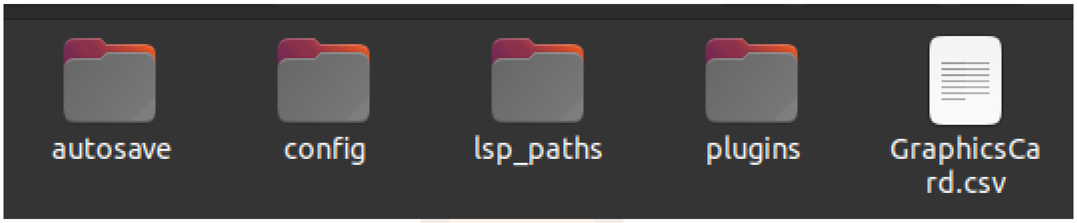
WRITING THE COLLECTED ORGANIZED HTML INTO A .CSV(EXCEL) FILE

```
f = open('GraphicsCard.csv','wb')
f.write('ProductName,Stars,Rating,CurrentPrice,ImageURL
\n'.encode())
for container in containers:
     #Finding product name
     product = container.findAll('a',{'class': 's1Q9rs'})
     ProductName = product[0].text.split(' Gr')[0]

     #Finding Stars
     star = container.find('div',{'class':'_3LWZlK'})
     try:
     Stars = star.text
     except:
     Stars = 0

     #Finding Ratings
     Rating = container.find('span',{'class':'_2_R_DZ'})
     try:
     Rating = Rating.text.replace('(','').replace(')','')
     except:
     Rating = 0

     #Finding Current Price
     CurrentPrice =
container.find('div',{'class':'_30jeq3'}).text.replace(',','')

     #Finding Image
     ImageURL = container.img
     ImageURL = (ImageURL.get('src'))

f.write(f"{ProductName},{Stars},{Rating},{CurrentPrice},{ImageUR
L}\n".encode())

f.close()

#EndOfCode
```

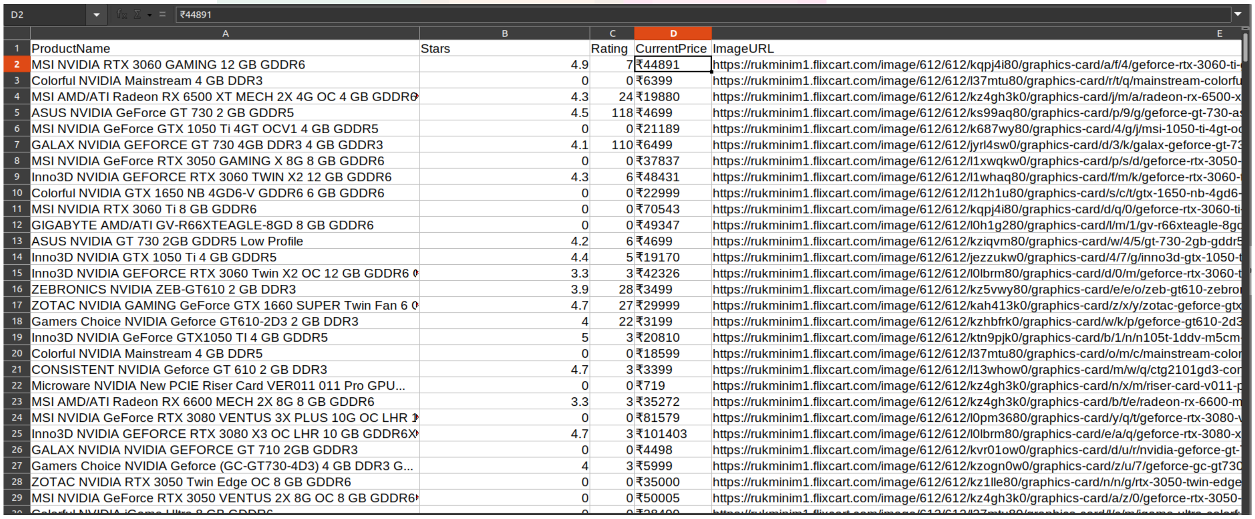


4.OUTPUT

The Output we get is a .csv file named - ‘GraphicsCard.csv’ :



On opening the .csv file:

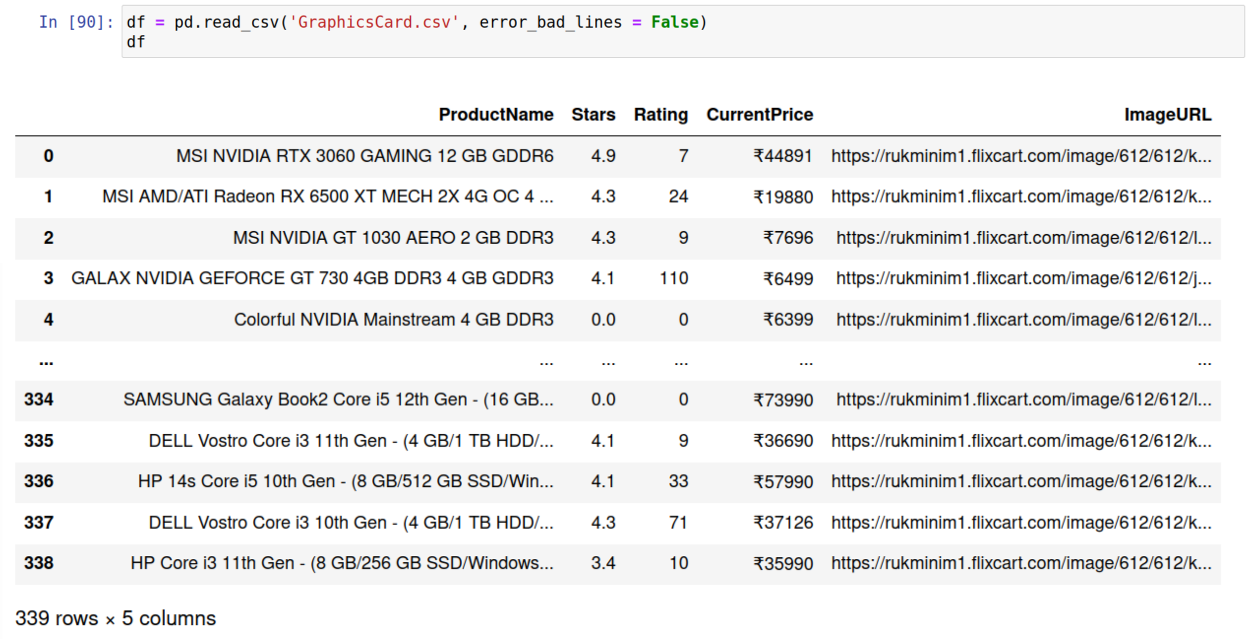


D2 ₹44891

	A	B	C	D	E
1	ProductName	Stars	Rating	CurrentPrice	ImageURL
2	MSI NVIDIA RTX 3060 GAMING 12 GB GDDR6	4.9	7	₹44891	https://rukminim1.flixcart.com/image/612/612/kqpj4i80/graphics-card/a/f/4/geforce-rtx-3060-ti-
3	Colorful NVIDIA Mainstream 4 GB DDR3	0	0	₹6399	https://rukminim1.flixcart.com/image/612/612/l37mtu80/graphics-card/r/t/q/mainstream-colorfu
4	MSI AMD/ATI Radeon RX 6500 XT MECH 2X 4G OC 4 GB GDDR6	4.3	24	₹19880	https://rukminim1.flixcart.com/image/612/612/kz4gh3k0/graphics-card/j/m/a/radeon-rx-6500-x
5	ASUS NVIDIA GeForce GT 730 2 GB GDDR5	4.5	118	₹4699	https://rukminim1.flixcart.com/image/612/612/ks99aq80/graphics-card/p/9/g/geforce-gt-730-a
6	MSI NVIDIA GeForce GTX 1050 Ti 4GT OCV1 4 GB GDDR5	0	0	₹21189	https://rukminim1.flixcart.com/image/612/612/k687wy80/graphics-card/4/g/j/msi-1050-ti-4gt-o
7	GALAX NVIDIA GEFORCE GT 730 4GB DDR3 4 GB GDDR3	4.1	110	₹6499	https://rukminim1.flixcart.com/image/612/612/jyrl4sw0/graphics-card/d/3/k/galax-geforce-gt-73
8	MSI NVIDIA GeForce RTX 3050 GAMING X 8G 8 GB GDDR6	0	0	₹37837	https://rukminim1.flixcart.com/image/612/612/l1xwqkw0/graphics-card/p/s/d/geforce-rtx-3050-
9	Inno3D NVIDIA GEFORCE RTX 3060 TWIN X2 12 GB GDDR6	4.3	6	₹48431	https://rukminim1.flixcart.com/image/612/612/l1whaq80/graphics-card/f/m/k/geforce-rtx-3060-
10	Colorful NVIDIA GTX 1650 NB 4GD6-V GDDR6 6 GB GDDR6	0	0	₹22999	https://rukminim1.flixcart.com/image/612/612/l12h1u80/graphics-card/s/c/t/gtx-1650-nb-4gd6-
11	MSI NVIDIA RTX 3060 Ti 8 GB GDDR6	0	0	₹70543	https://rukminim1.flixcart.com/image/612/612/kqpj4i80/graphics-card/d/q/0/geforce-rtx-3060-ti
12	GIGABYTE AMD/ATI GV-R66XTEAGLE-8GD 8 GB GDDR6	0	0	₹49347	https://rukminim1.flixcart.com/image/612/612/l0h1g280/graphics-card/l/m/1/gv-r66xteagle-8g
13	ASUS NVIDIA GT 730 2GB GDDR5 Low Profile	4.2	6	₹4699	https://rukminim1.flixcart.com/image/612/612/kziqvm80/graphics-card/w/4/5/gt-730-2gb-gddr
14	Inno3D NVIDIA GTX 1050 Ti 4 GB GDDR5	4.4	5	₹19170	https://rukminim1.flixcart.com/image/612/612/jezzukw0/graphics-card/4/7/g/inno3d-gtx-1050-
15	Inno3D NVIDIA GEFORCE RTX 3060 Twin X2 OC 12 GB GDDR6	3.3	3	₹42326	https://rukminim1.flixcart.com/image/612/612/l0lbrm80/graphics-card/d/0/m/geforce-rtx-3060-
16	ZEBRONICS NVIDIA ZEB-GT610 2 GB DDR3	3.9	28	₹3499	https://rukminim1.flixcart.com/image/612/612/kz5vwy80/graphics-card/e/e/o/zeb-gt610-zebro
17	ZOTAC NVIDIA GAMING GeForce GTX 1660 SUPER Twin Fan 6	4.7	27	₹29999	https://rukminim1.flixcart.com/image/612/612/kah413k0/graphics-card/z/x/y/zotac-geforce-gtx
18	Gamers Choice NVIDIA Geforce GT610-2D3 2 GB DDR3	4	22	₹3199	https://rukminim1.flixcart.com/image/612/612/kzhbfrk0/graphics-card/w/k/p/geforce-gt610-2d3
19	Inno3D NVIDIA GeForce GTX1050 TI 4 GB GDDR5	5	3	₹20810	https://rukminim1.flixcart.com/image/612/612/ktn9pjk0/graphics-card/b/1/n/n105t-1ddv-m5cm
20	Colorful NVIDIA Mainstream 4 GB DDR5	0	0	₹18599	https://rukminim1.flixcart.com/image/612/612/l37mtu80/graphics-card/o/m/c/mainstream-color
21	CONSISTENT NVIDIA Geforce GT 610 2 GB DDR3	4.7	3	₹3399	https://rukminim1.flixcart.com/image/612/612/l13whow0/graphics-card/m/w/q/ctg2101gd3-co
22	Microware NVIDIA New PCIE Riser Card VER011 011 Pro GPU...	0	0	₹719	https://rukminim1.flixcart.com/image/612/612/kz4gh3k0/graphics-card/n/x/m/riser-card-v011-
23	MSI AMD/ATI Radeon RX 6600 MECH 2X 8G 8 GB GDDR6	3.3	3	₹35272	https://rukminim1.flixcart.com/image/612/612/kz4gh3k0/graphics-card/b/t/e/radeon-rx-6600-m
24	MSI NVIDIA GeForce RTX 3080 VENTUS 3X PLUS 10G OC LHR 1	0	0	₹81579	https://rukminim1.flixcart.com/image/612/612/l0pm3680/graphics-card/y/q/t/geforce-rtx-3080-
25	Inno3D NVIDIA GEFORCE RTX 3080 X3 OC LHR 10 GB GDDR6X	4.7	3	₹101403	https://rukminim1.flixcart.com/image/612/612/l0lbrm80/graphics-card/e/a/q/geforce-rtx-3080-x
26	GALAX NVIDIA NVIDIA GEFORCE GT 710 2GB GDDR3	0	0	₹4498	https://rukminim1.flixcart.com/image/612/612/kvr01ow0/graphics-card/d/u/r/nvidia-geforce-gt-
27	Gamers Choice NVIDIA Geforce (GC-GT730-4D3) 4 GB DDR3 G...	4	3	₹5999	https://rukminim1.flixcart.com/image/612/612/kzogn0w0/graphics-card/z/u/7/geforce-gc-gt730
28	ZOTAC NVIDIA RTX 3050 Twin Edge OC 8 GB GDDR6	0	0	₹35000	https://rukminim1.flixcart.com/image/612/612/kz1lle80/graphics-card/n/n/g/rtx-3050-twin-edge
29	MSI NVIDIA GeForce RTX 3050 VENTUS 2X 8G OC 8 GB GDDR6	0	0	₹50005	https://rukminim1.flixcart.com/image/612/612/kz4gh3k0/graphics-card/a/z/0/geforce-rtx-3050-

We get all the Names, Stars, No. of Ratings & the Image URL of the search results. The scraper has collected 401 search results with 10 pages of the website.

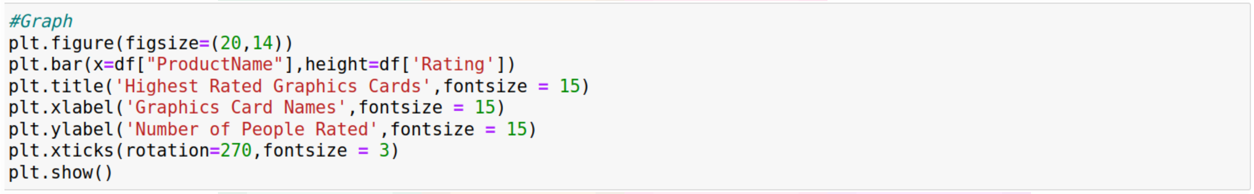
We can further analyze the results by using numpy, pandas & matplotlib as follows:



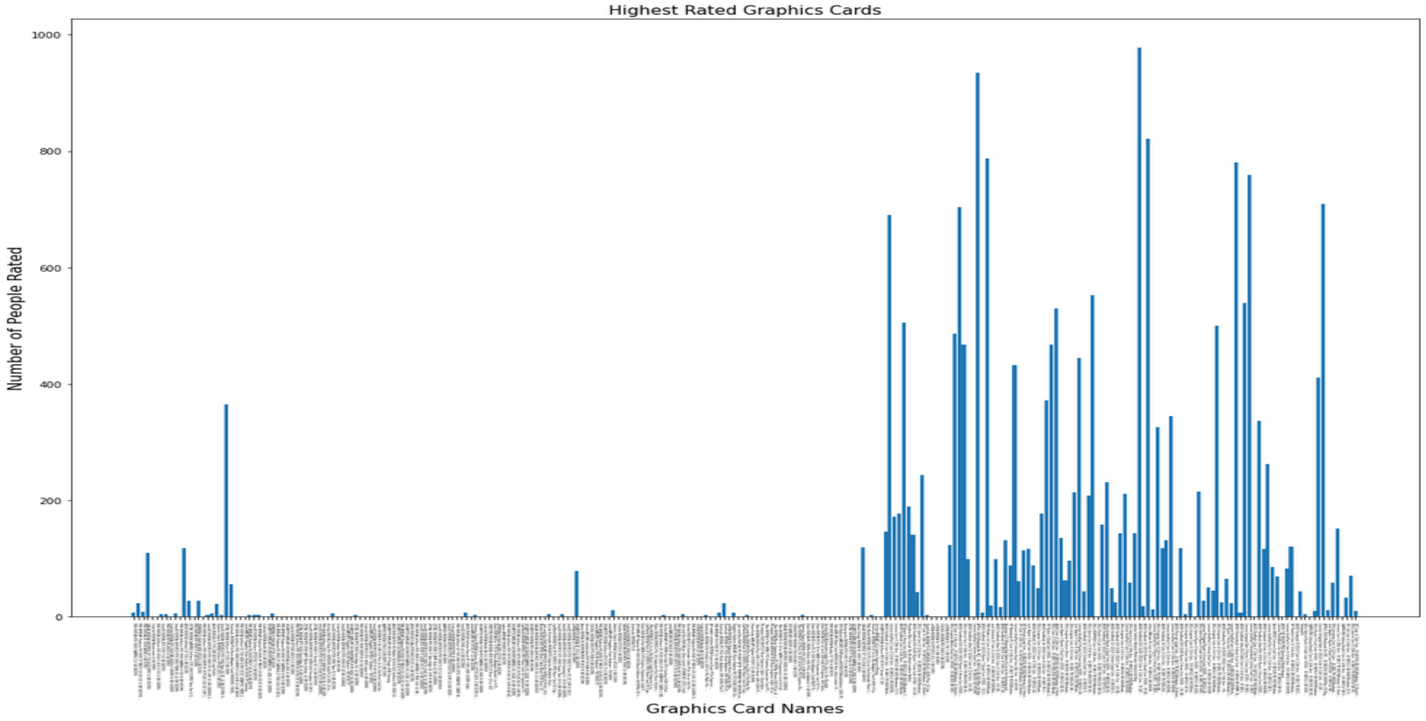
```
In [90]: df = pd.read_csv('GraphicsCard.csv', error_bad_lines = False)
         df
```

	ProductName	Stars	Rating	CurrentPrice	ImageURL
0	MSI NVIDIA RTX 3060 GAMING 12 GB GDDR6	4.9	7	₹44891	https://rukminim1.flixcart.com/image/612/612/k...
1	MSI AMD/ATI Radeon RX 6500 XT MECH 2X 4G OC 4 ...	4.3	24	₹19880	https://rukminim1.flixcart.com/image/612/612/k...
2	MSI NVIDIA GT 1030 AERO 2 GB DDR3	4.3	9	₹7696	https://rukminim1.flixcart.com/image/612/612/l...
3	GALAX NVIDIA GEFORCE GT 730 4GB DDR3 4 GB GDDR3	4.1	110	₹6499	https://rukminim1.flixcart.com/image/612/612/j...
4	Colorful NVIDIA Mainstream 4 GB DDR3	0.0	0	₹6399	https://rukminim1.flixcart.com/image/612/612/l...
...	...	...	...	...	...
334	SAMSUNG Galaxy Book2 Core i5 12th Gen - (16 GB...	0.0	0	₹73990	https://rukminim1.flixcart.com/image/612/612/l...
335	DELL Vostro Core i3 11th Gen - (4 GB/1 TB HDD/...	4.1	9	₹36690	https://rukminim1.flixcart.com/image/612/612/k...
336	HP 14s Core i5 10th Gen - (8 GB/512 GB SSD/Win...	4.1	33	₹57990	https://rukminim1.flixcart.com/image/612/612/k...
337	DELL Vostro Core i3 10th Gen - (4 GB/1 TB HDD/...	4.3	71	₹37126	https://rukminim1.flixcart.com/image/612/612/k...
338	HP Core i3 11th Gen - (8 GB/256 GB SSD/Windows...	3.4	10	₹35990	https://rukminim1.flixcart.com/image/612/612/k...

339 rows × 5 columns



```
#Graph
plt.figure(figsize=(20,14))
plt.bar(x=df["ProductName"],height=df['Rating'])
plt.title('Highest Rated Graphics Cards',fontsize = 15)
plt.xlabel('Graphics Card Names',fontsize = 15)
plt.ylabel('Number of People Rated',fontsize = 15)
plt.xticks(rotation=270,fontsize = 3)
plt.show()
```

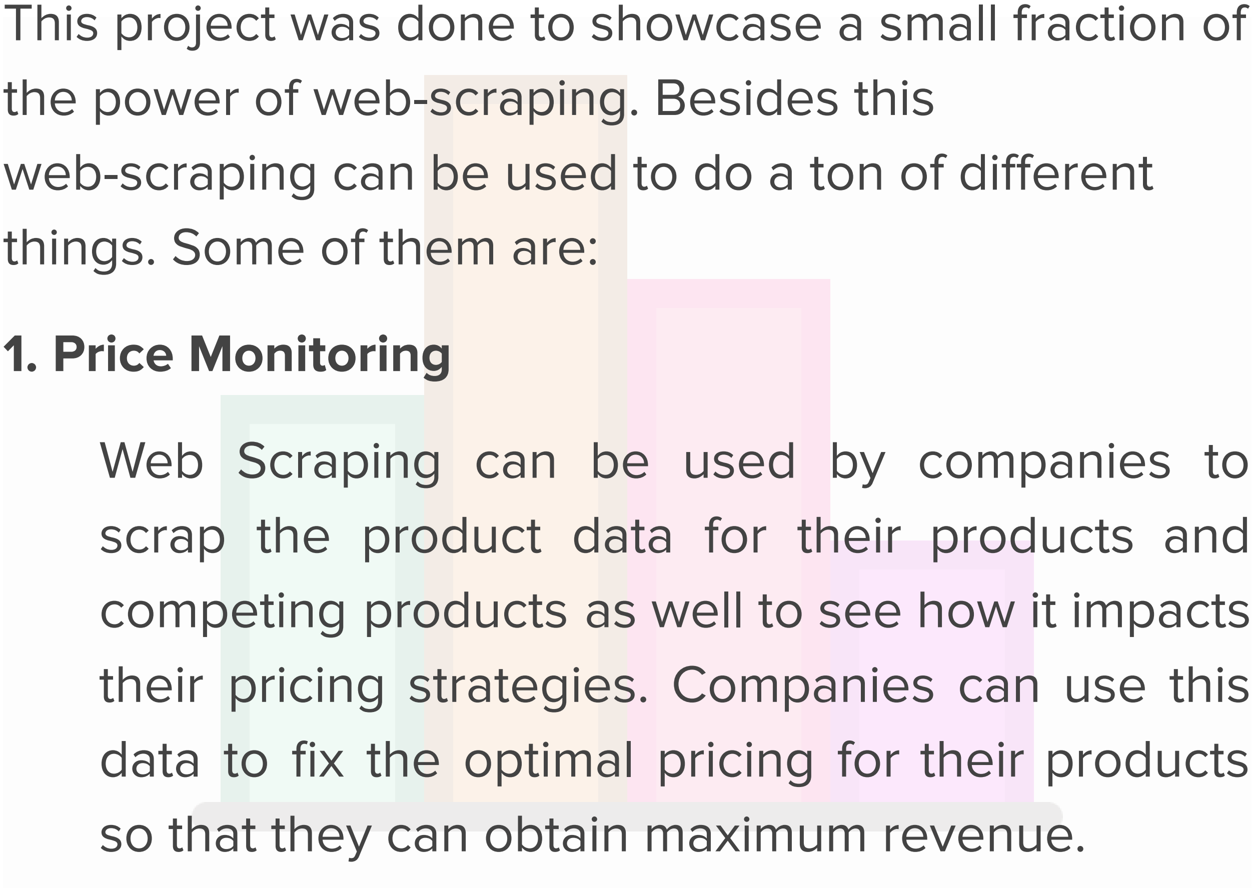


5.FUTURE SCOPE OF THIS PROJECT

This project was done to showcase a small fraction of the power of web-scraping. Besides this web-scraping can be used to do a ton of different things. Some of them are:

1. Price Monitoring

Web Scraping can be used by companies to scrap the product data for their products and competing products as well to see how it impacts their pricing strategies. Companies can use this data to fix the optimal pricing for their products so that they can obtain maximum revenue.

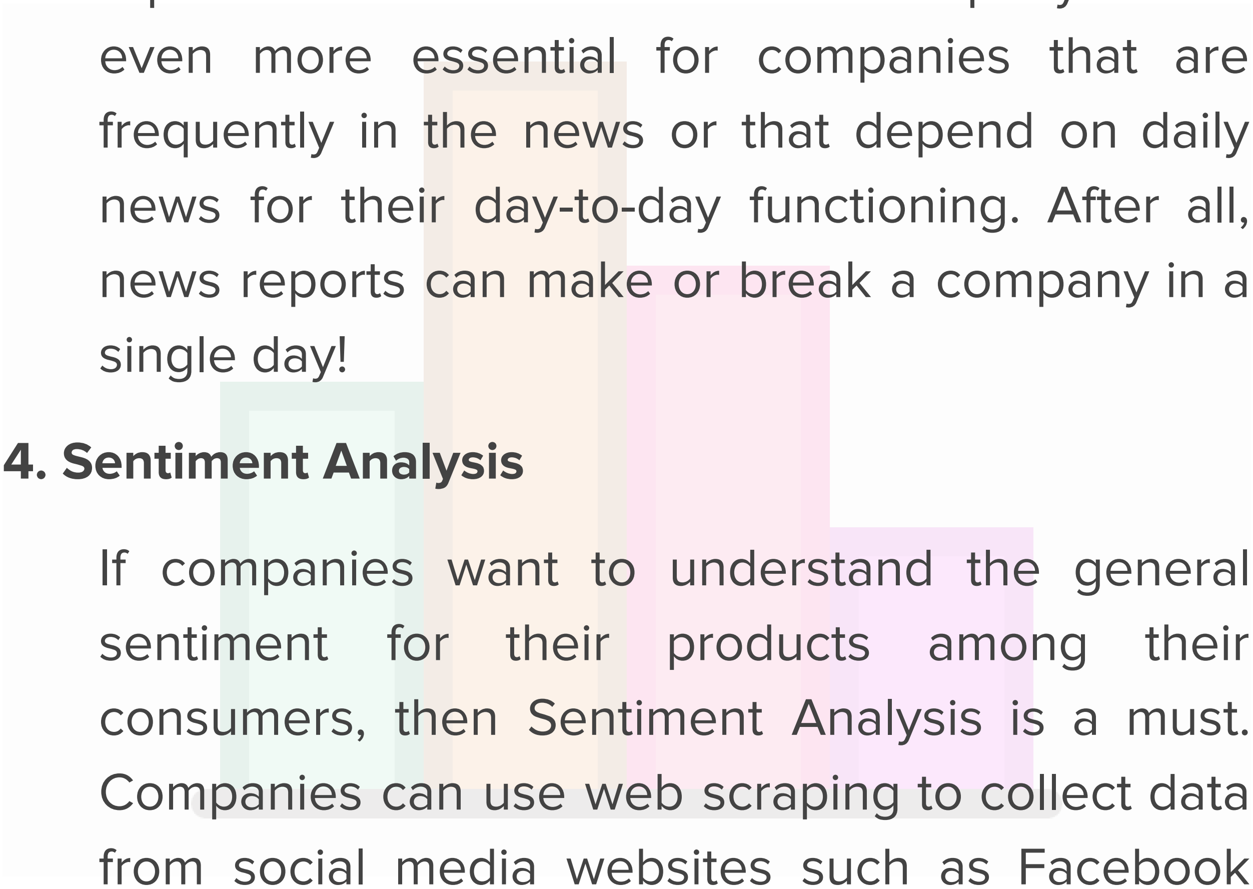


2. Market Research

Web scraping can be used for market research by companies. High-quality web scraped data obtained in large volumes can be very helpful for companies in analyzing consumer trends and understanding which direction the company should move in the future.

3. News Monitoring

Web scraping news sites can provide detailed reports on the current news to a company. This is even more essential for companies that are frequently in the news or that depend on daily news for their day-to-day functioning. After all, news reports can make or break a company in a single day!

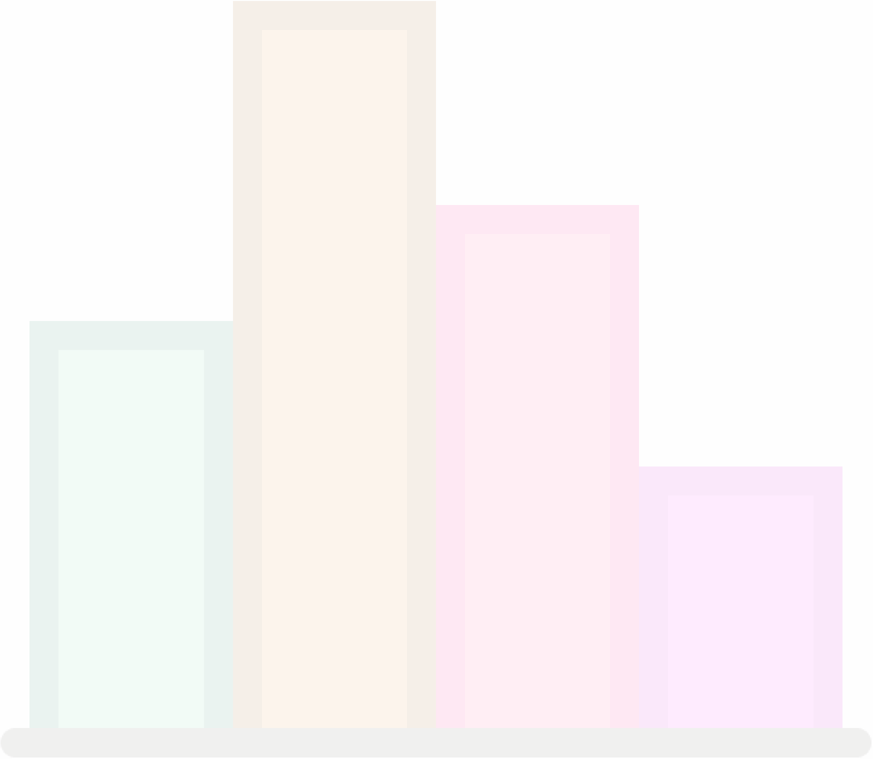


4. Sentiment Analysis

If companies want to understand the general sentiment for their products among their consumers, then Sentiment Analysis is a must. Companies can use web scraping to collect data from social media websites such as Facebook and Twitter as to what the general sentiment about their products is. This will help them in creating products that people desire and moving ahead of their competition.

5. Email Marketing

Companies can also use Web scraping for email marketing. They can collect Email ID’s from various sites using web scraping and then send bulk promotional and marketing Emails to all the people owning these Email ID’s.



6.BIBLIOGRAPHY

https://www.geeksforgeeks.org/what-is-web-scraping-and-how-to-use-it/

https://www.flipkart.com/search?q=graphics+card

https://en.wikipedia.org/wiki/Web_scraping

https://www.youtube.com/watch?v=dQw4w9WgXcQ

https://www.youtube.com/watch?v=yscomtiwmsk&list=PLPBPWin4aWklCQu9q2zLQmspH_GQtVDTe