**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**



| **S.no** | **Contents** |
| --- | --- |
| **1.** | [Introduction](#_ew6p7vozcmyw) |
| **2.** | [Project Setup](#_xqz99z7w5b9f) |
| **3.** | [Source Code](#_5yvphdbqpjap) |
| **4.** | [Output](#_800gh77nid7r) |
| **5.** | [Future Scope](#_zc8rb857lf8p) |
| **6.** | [Bibliography](#_4v4go65mx3mm) |

# **1.INTRODUCTION**

**What is Web Scraping?**

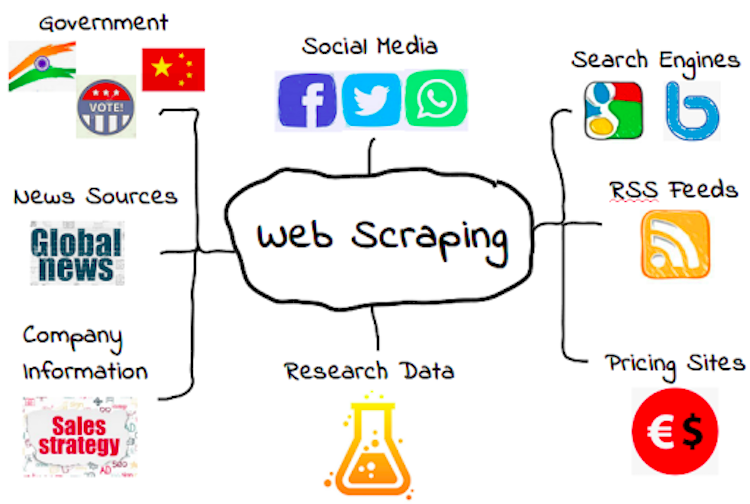
Web scraping is an automatic method to obtain large amounts of data from websites. Most of this data is unstructured data in an HTML format which is then converted into structured data in a spreadsheet or a database so that it can be used in various applications. There are many different ways to perform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

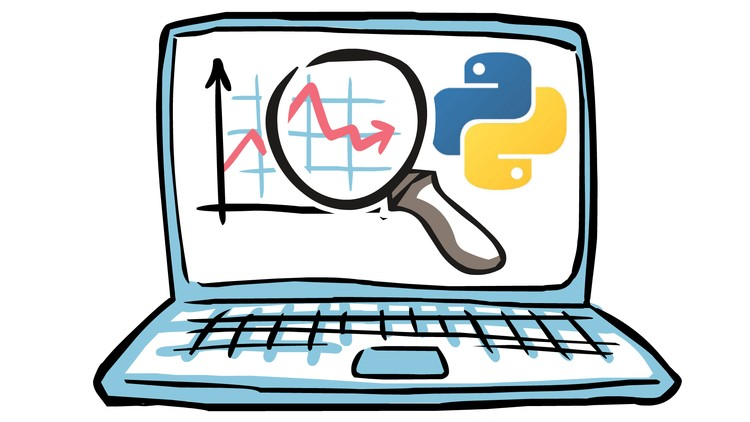
Python seems to be in fashion these days! It is the most popular language for web scraping as it can handle most of the processes easily. It also has a variety of libraries that were created specifically for Web Scraping. **Scrapy** is a very popular open-source web crawling framework that is written 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:**



1. **bs4:**



**Optional:**

1. **pandas:**

****

1. **numpy:**

****

1. **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

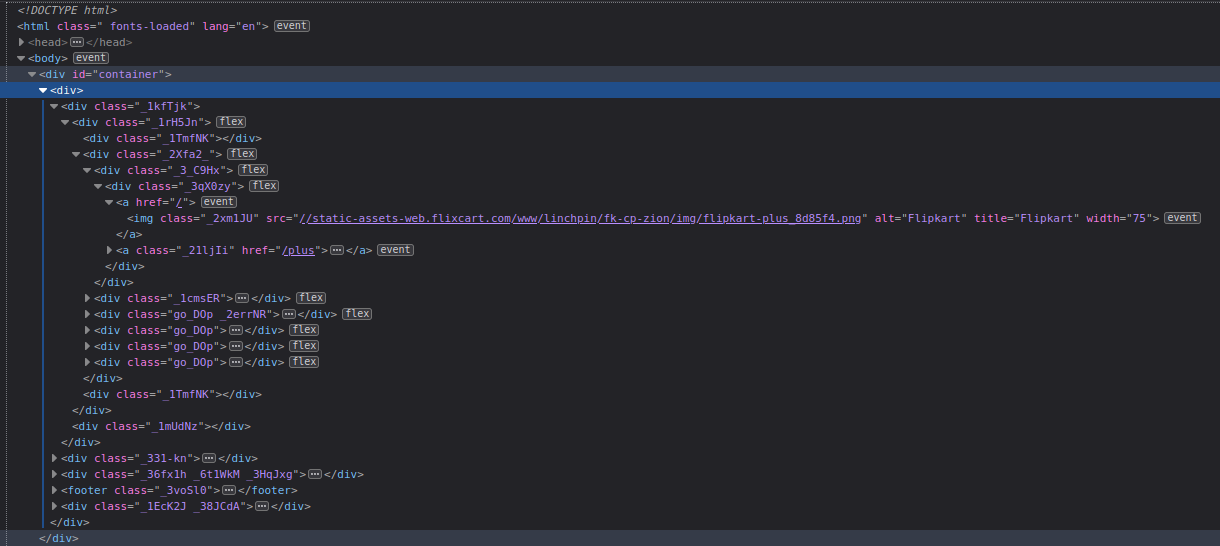
import matplotlib.pyplot as plt

**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)

**\*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):

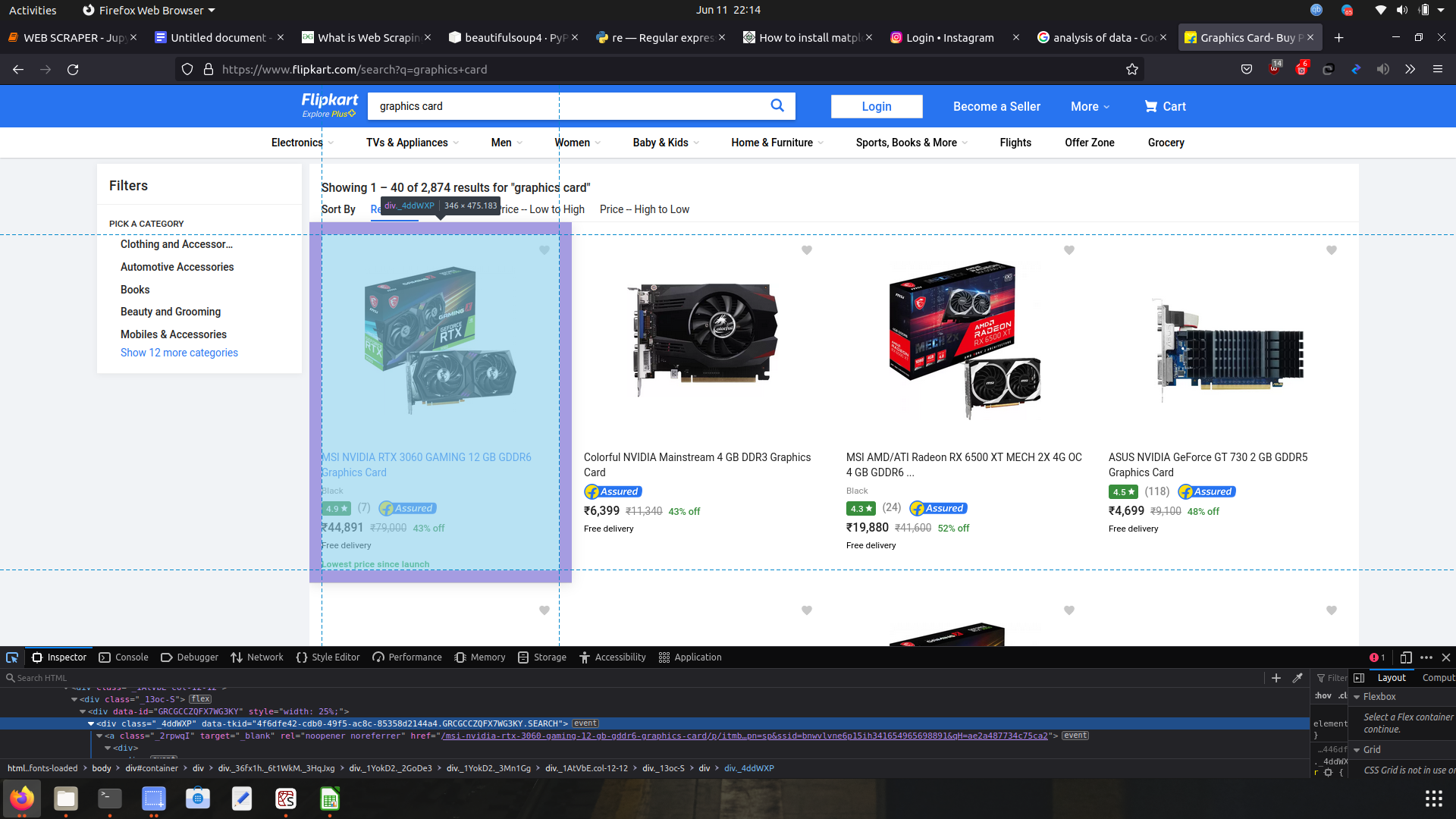
link = domain+str(i)

page\_data = urlopen(link)

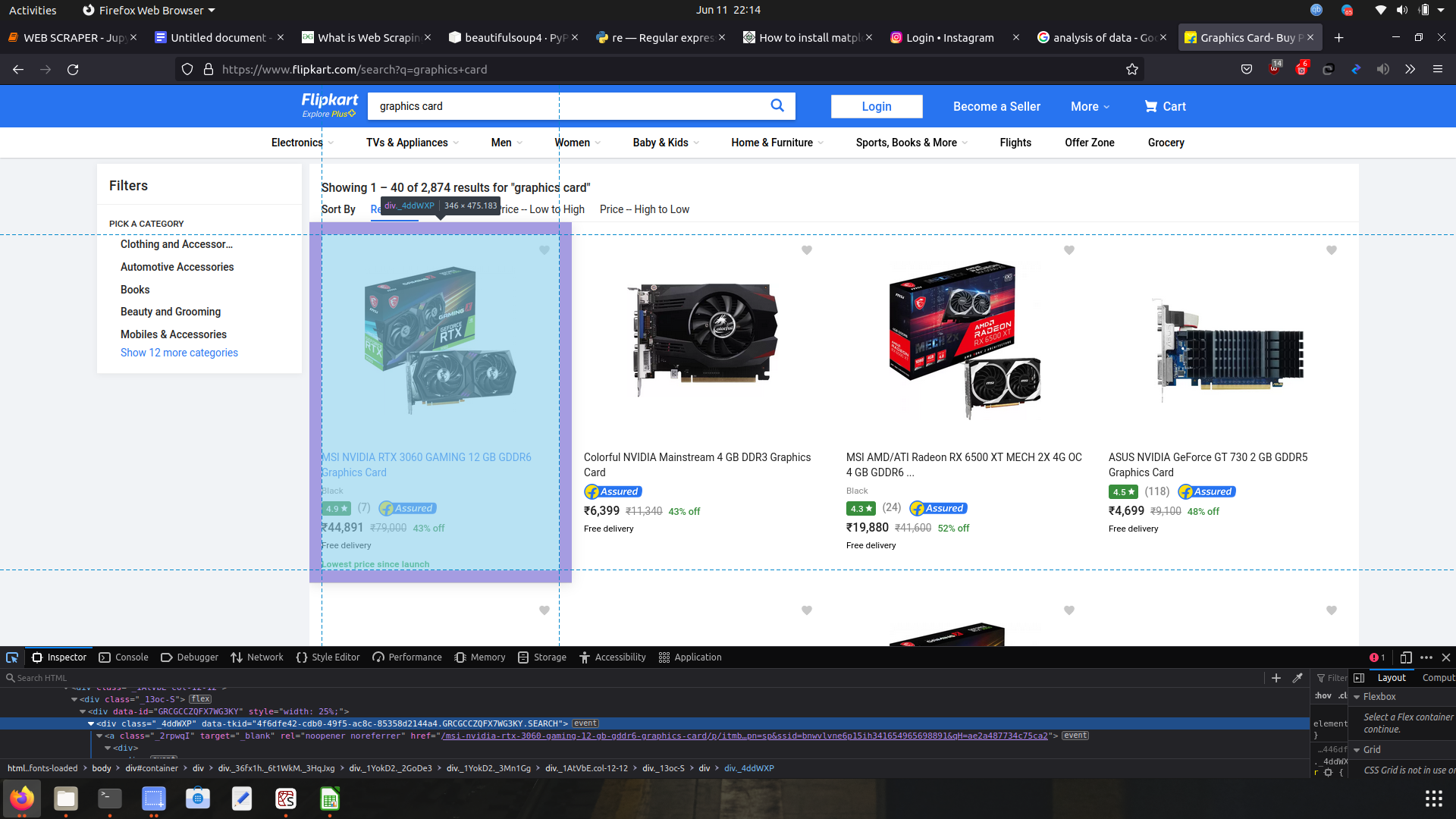
webdata1 = page\_data.read()

webdata += webdata1

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



**\*IMPORTANT HTML\***



**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},{ImageURL}\n".encode())



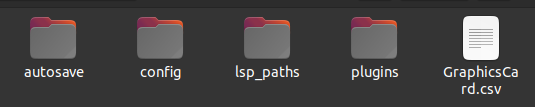
f.close()

#EndOfCode

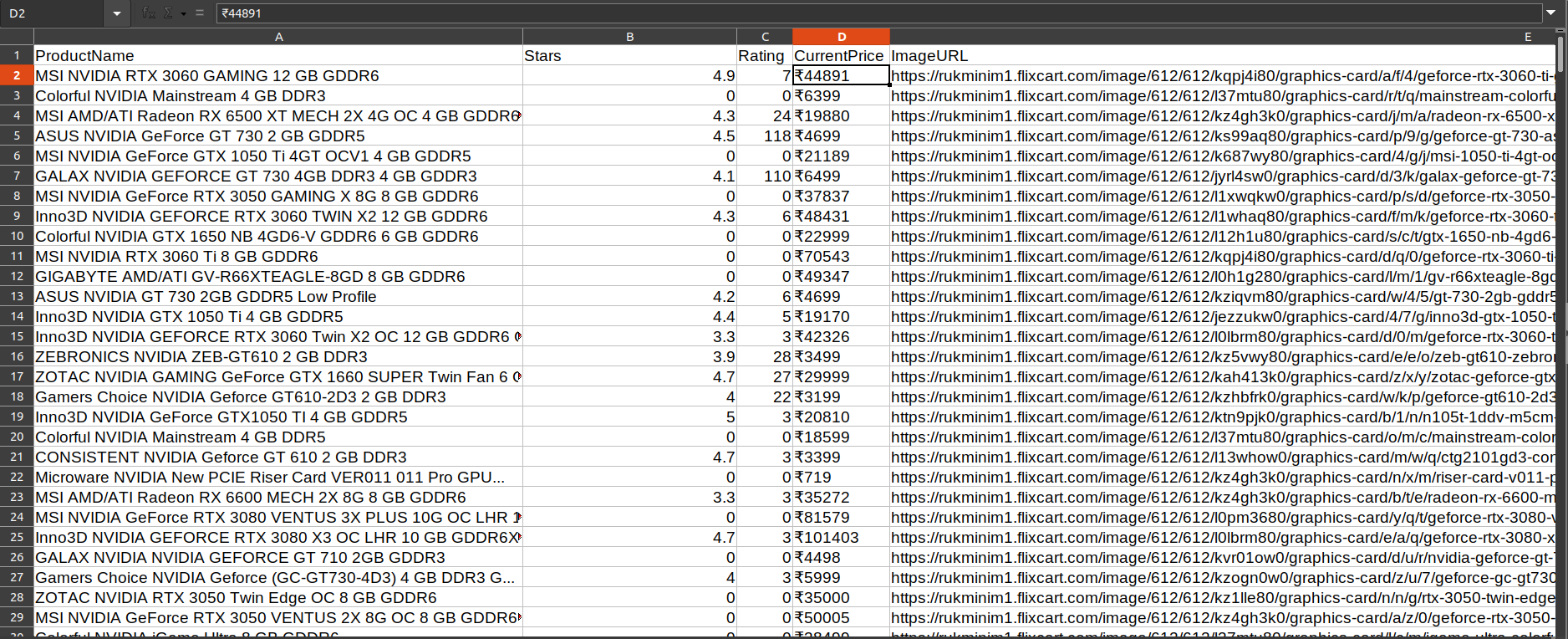


# **4.OUTPUT**

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

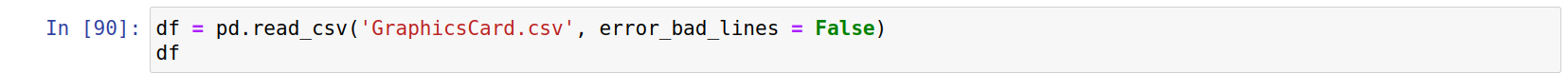


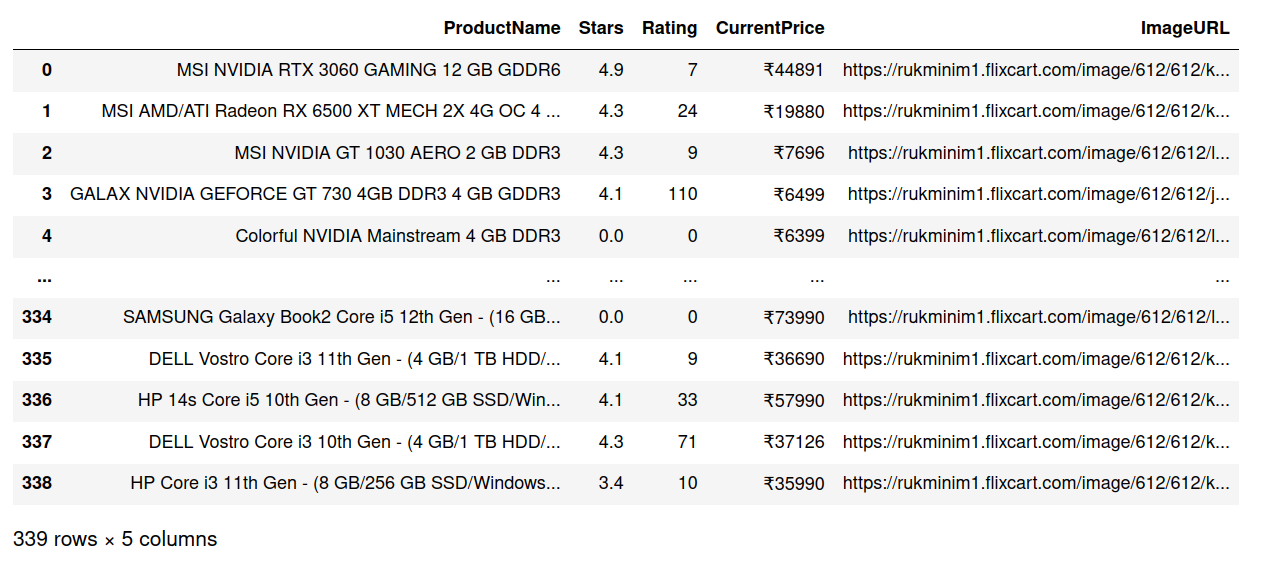
On opening the .csv file:

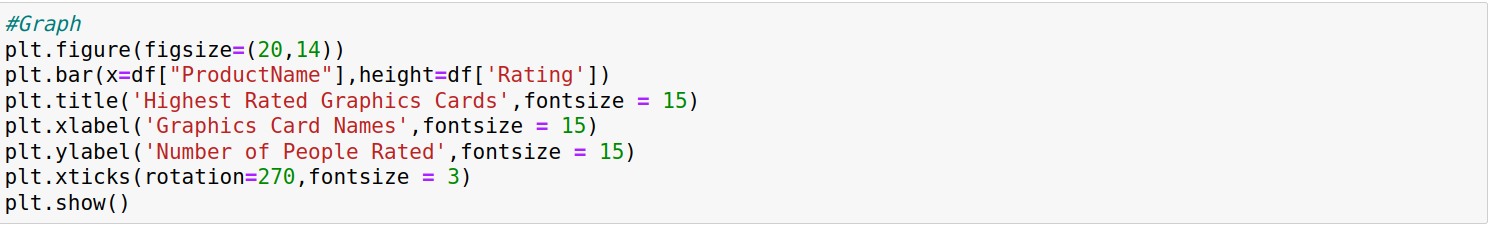


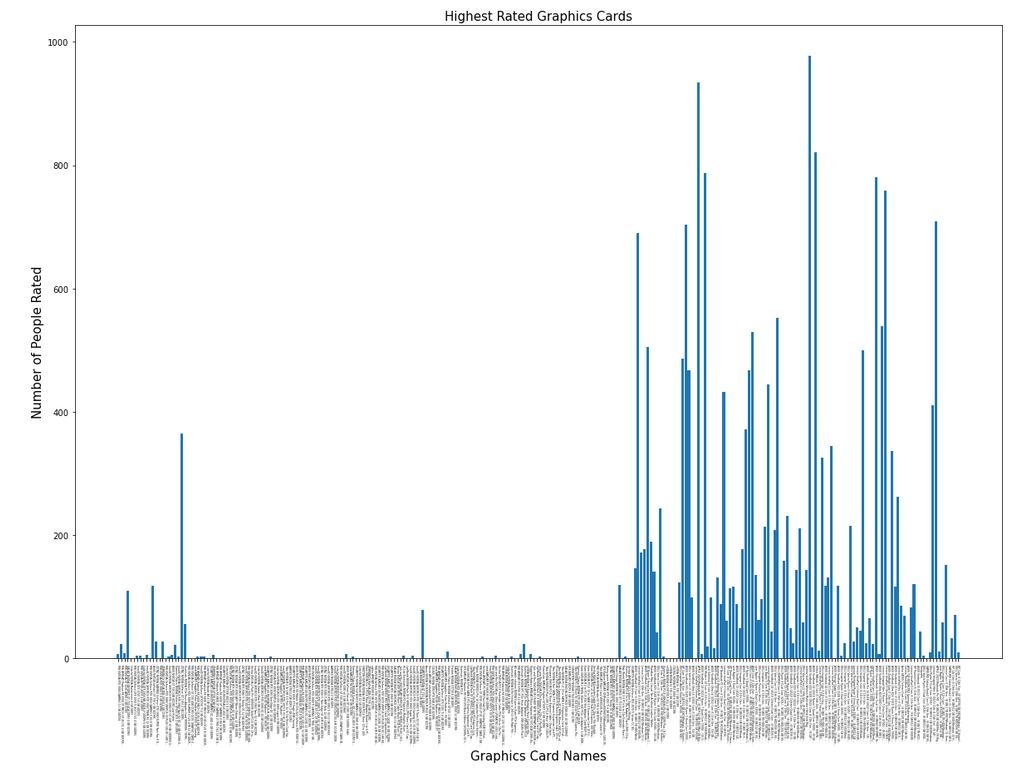
We get all the Names, Stars, No. of Ratings & the Image URL of the search results. The scraper has collected 401 search results with 10 pages of the website.

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









# **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>