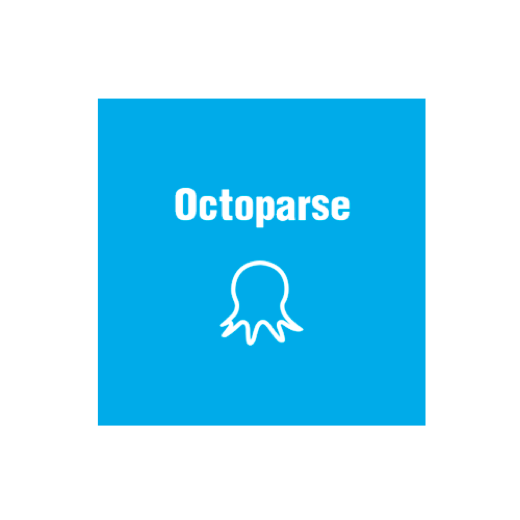
****

### Web Scraping

#### Q1. What is We

Web Scraping

#### Q1. What is Web Scraping?

#### ****Web Scraping**** is the process through which we extract data from a website, and save it in a form which is easy to read, to understand and to work on.

#### When we say 'Easy to work on', we mean to say that the data thus extracted can be used to get a lot of useful insights and answer a lot of questions, finding answers to which would not be such an easy task, if we did not have that data stored with us in a simple and sorted manner, i.e. generally in a CSV File, an Excel File or a Database.

#### web scraping work.png

Q2. How does web scraping work?

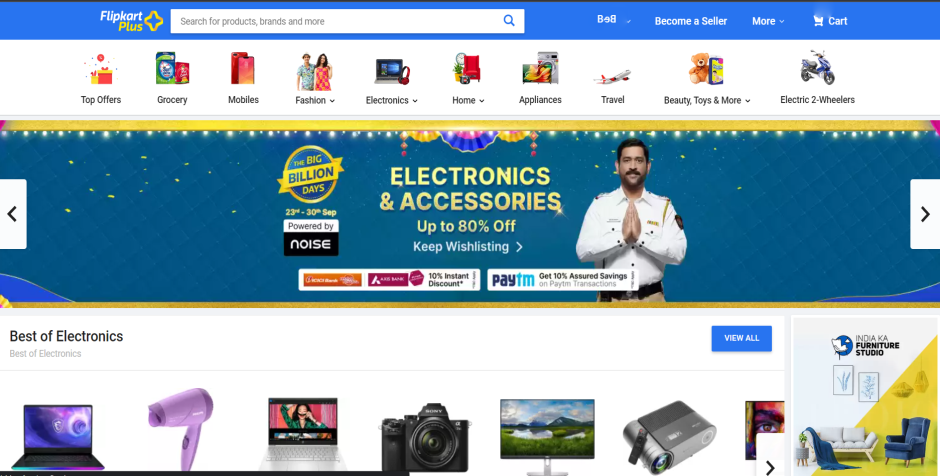
To understand web scraping, it’s important to first understand that web pages are built with text-based mark-up languages – the most common being HTML.

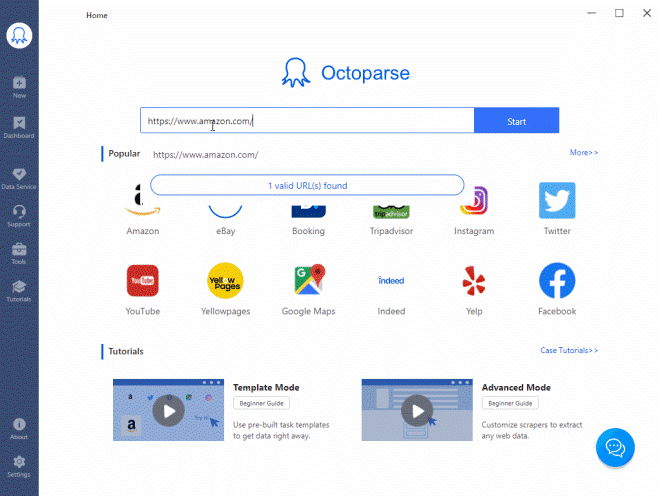
A mark-up language defines the structure of a website’s content. Since there are universal components and tags of mark-up languages, this makes it much easier for web scrapers to pull all the information that it needs. Once the HTML is parsed, the scraper then extracts the necessary data and stores it.  
**Note** : Not all websites allow Web Scraping, especially when personal information of the users is involved, so we should always ensure that we do not explore too much, and don't get our hands on information which might belong to someone else. Websites generally have protections at place, and they would block our access to the website if they see us scraping a large amount of data from their website.

Although web scraping can be done manually, in most cases, you might be better off using an automated tool. After all, these are usually faster and less expensive than scraping data manually.

Some of the automatic web scrapers are Octoparse etc.

In this Project we are using Automatic Web scraper –Octoparse.

**About Flipkart and Octoparse**

****

**Flipkart** is an Indian [e-commerce](https://en.wikipedia.org/wiki/E-commerce) company, headquartered in [Bangalore](https://en.wikipedia.org/wiki/Bangalore), and incorporated in [Singapore](https://en.wikipedia.org/wiki/Singapore) as a [private limited company](https://en.wikipedia.org/wiki/Private_company_limited_by_shares). The company initially focused on online book sales before expanding into other product categories such as [consumer electronics](https://en.wikipedia.org/wiki/Consumer_electronics), fashion, home essentials, groceries, and lifestyle products.

Octoparse is a cloud-based web data extraction solution that helps users extract relevant information from various types of websites. It enables users from a variety of industries to scrape unstructured data and save it in different formats including Excel, plain text and HTML.

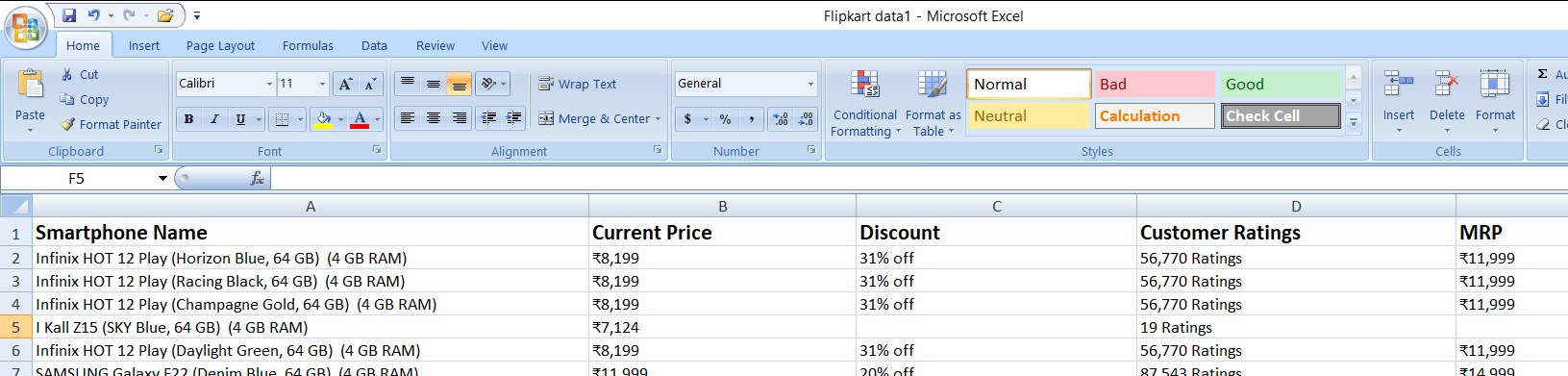
**Project Idea**

In this project we will parse through the flipkart website for smartphones category that are available on flipkart for purchase. We will take the price information along with the discount which is available at that time ,what is the colour and which brand mobile was it. We will also mtake info about user Rating that how many review does the phone have and the Mrp of the particular smartphone.

We will retrieve information from the page ‘**Smartphones**’ using web scraping: a process of extracting information from a website.Automaticaly Web scraping isn’t magic, and yet some readers may grab information on a daily basis.

**Project Goal**

The project goal is to build a web scraper that withdraws all desirable information and assemble them into a single CSV. The format of the output CSV or .xlxs file is shown below:



**Project Steps**

Here is an outline of the steps we follow.

1.Search for smartphones in flipkart.

2.Copy the link and paste it into the Octoparse Application.

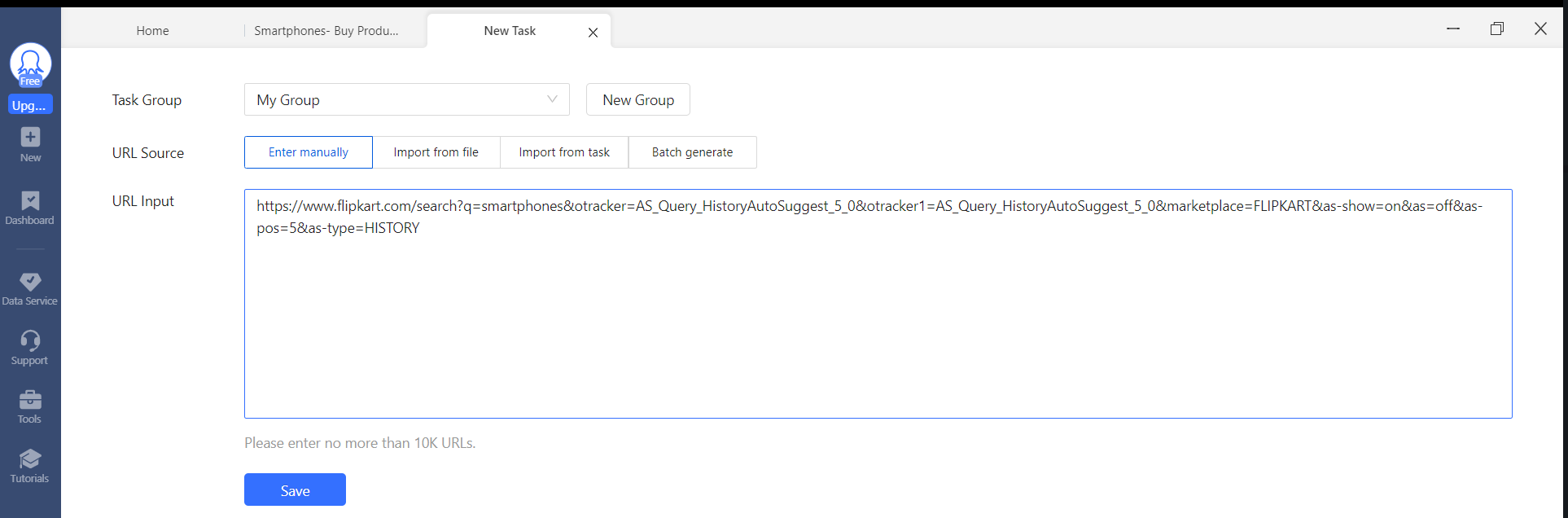
3.Select the data one by one to extract.

4.Create Pagination to extract all the data .And start loop url.

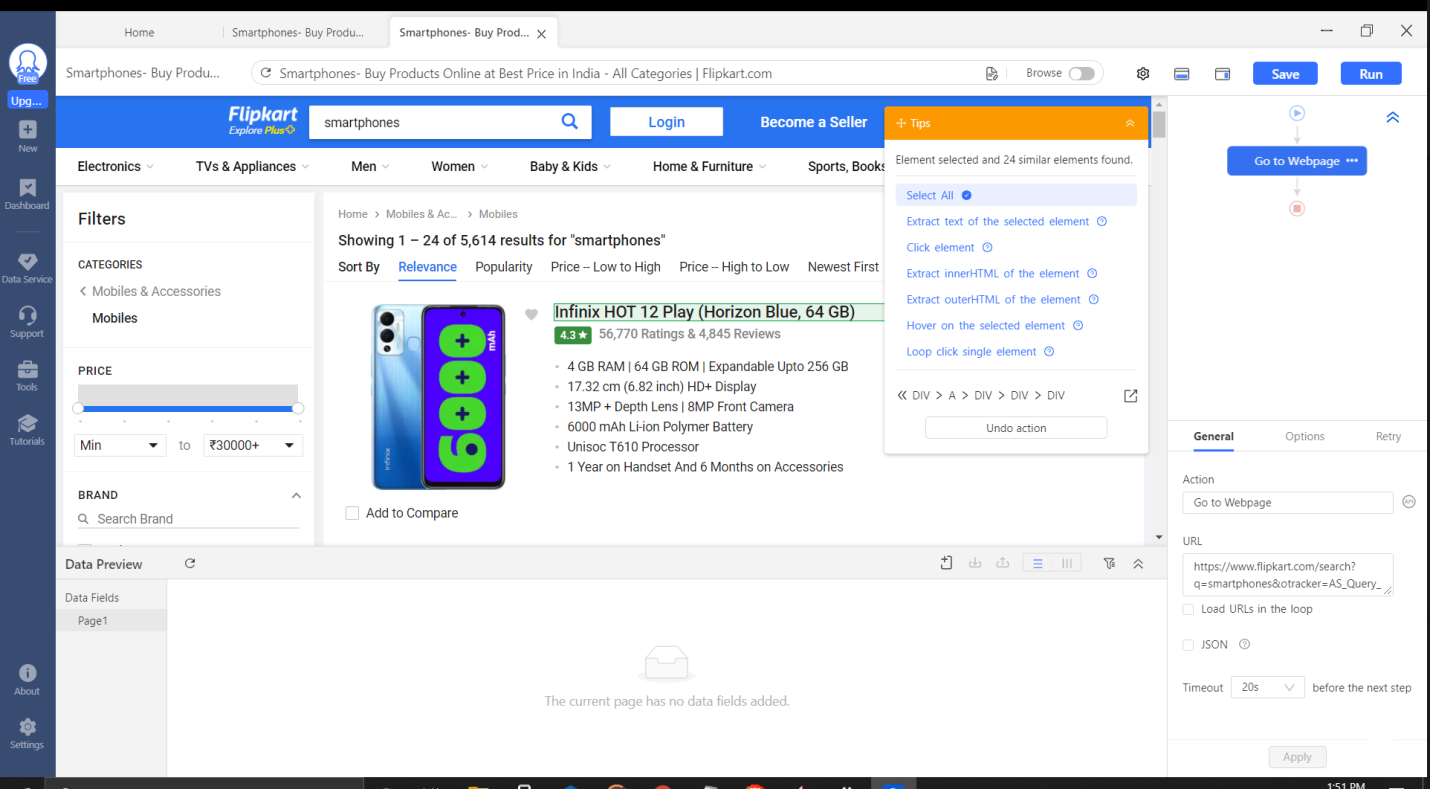
5.Save and Run ,It will start extracting the data.

**Let’s Begin**

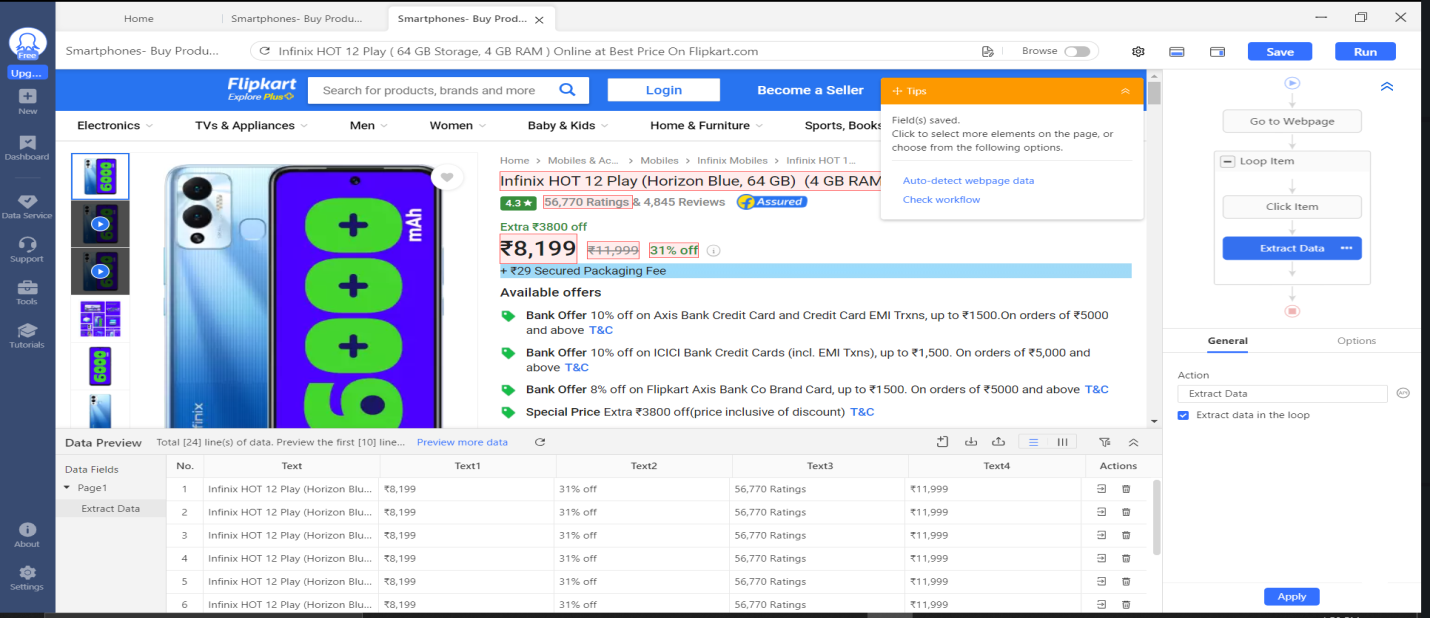
1.Download Octoparse from official website.Signup for account and open the Dashboard. Open Flipkart and search for smartphones copy the link and paste it to the URL input in the Octoparse dashboard and click save.



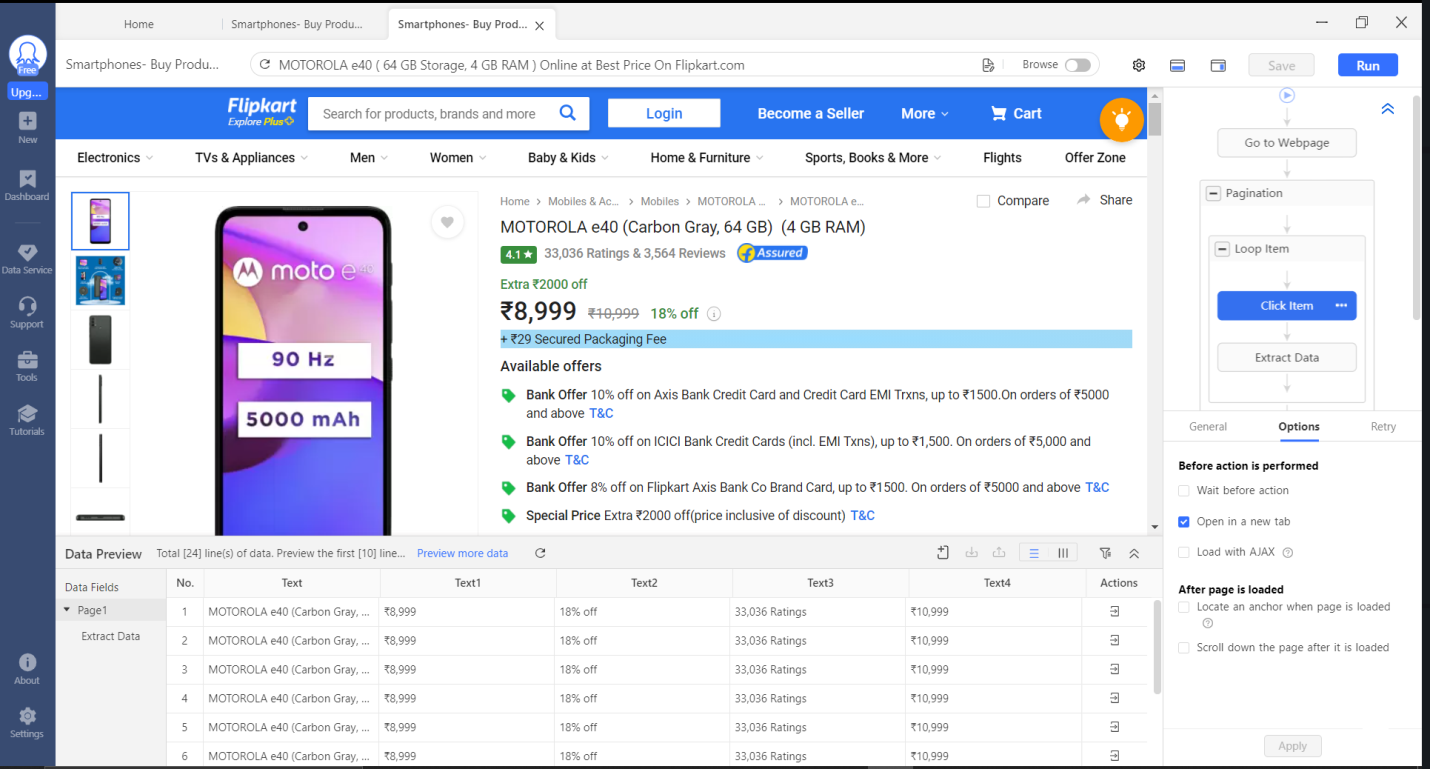
2. The same page will open in the octoparse it will start extracting automatically –cancel it select manually that which part you want to extract click on that a small window a popup will appear in the right corner in that click select all. It will select all the name of the smartphones in the page.



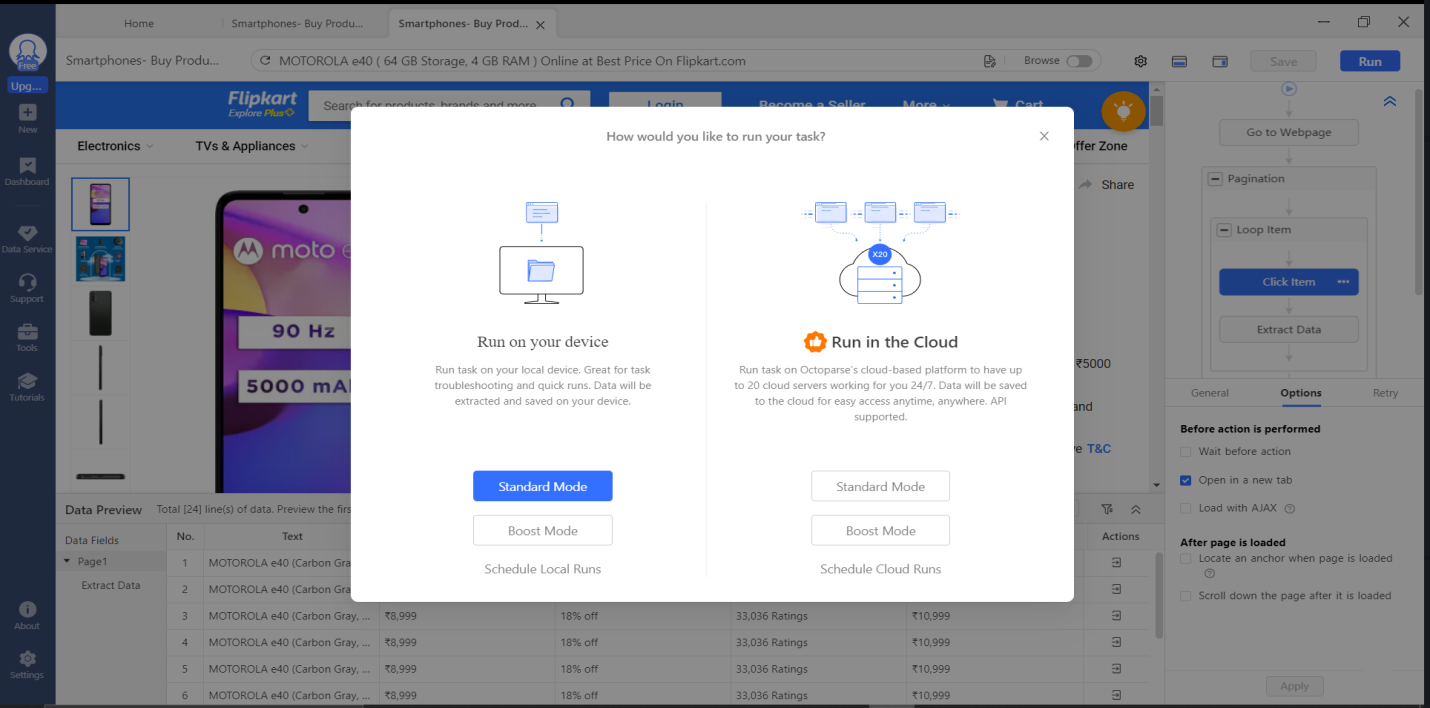
3. Select all the desired data you want to extract .From this page we are extracting the name of the smartphone, Discount available, price ,customer ratings that how many of them purchased that particular smartphone, and the Mrp that what is the mentioned price in the box of the smartphone. After that in the right side we can see the option extract data in loop click on that, the overview of the data will be shown below .



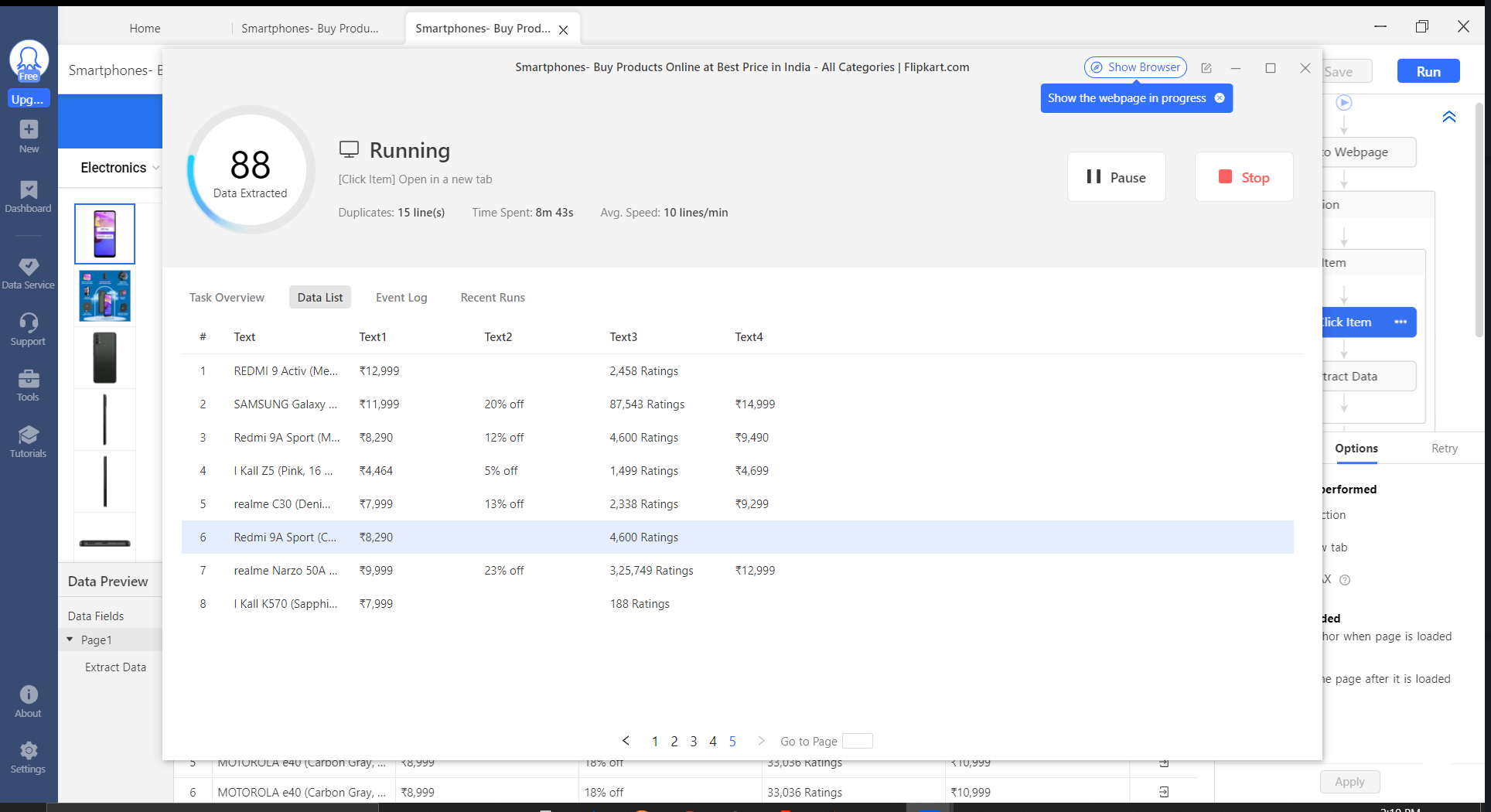
4. Click on save button in the right hand side in the task bar. Below you can see the preview of the data that will be extracted. Check the data that it is in correct order ,if not select the data according step by step ex-name>current price>Discount>Ratings>mrp --- .



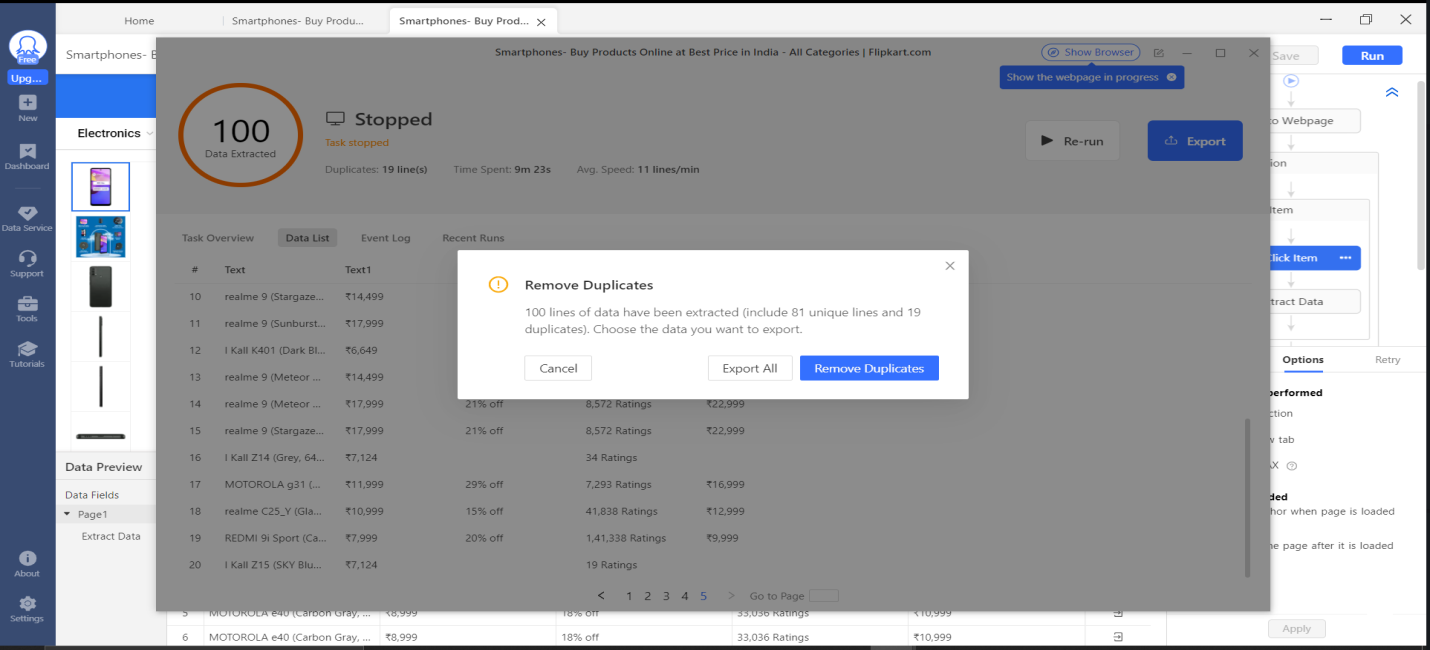
5. Click on the run button in the right side of the taskbar, A popup will show that how would you like to run click on run on this Device.



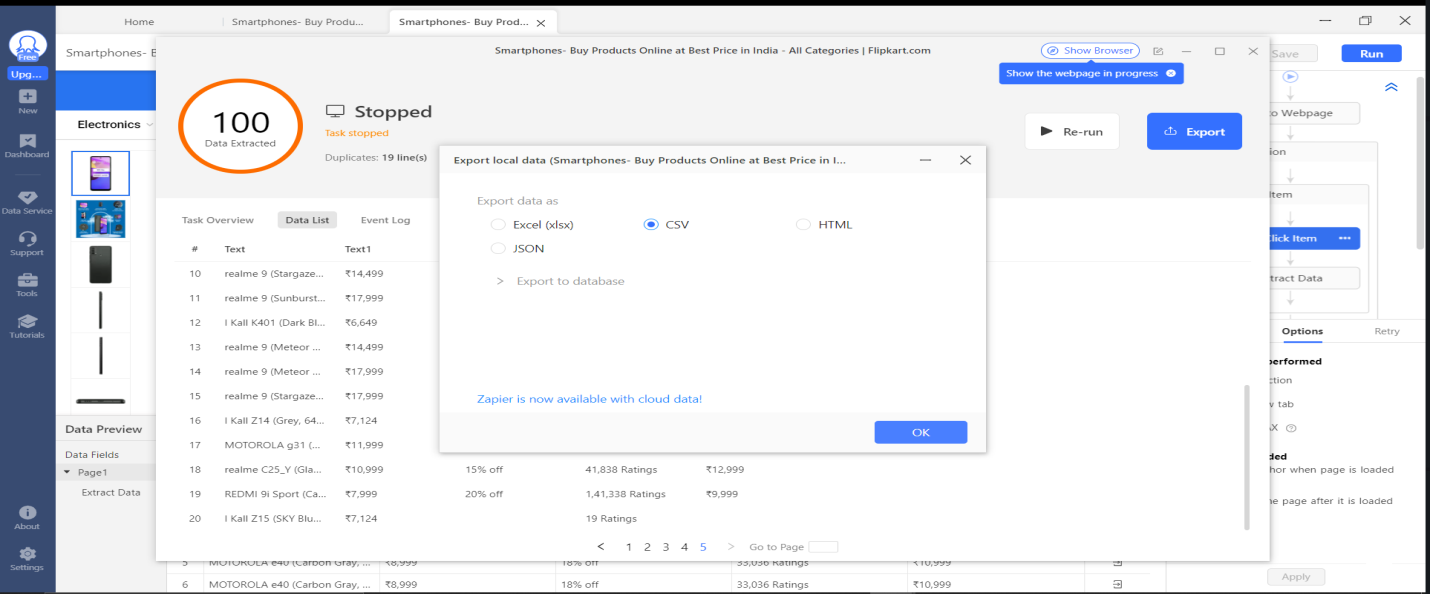
6. New window will open and now data will start extracting it will show all the progress , how much data extracted every info in the window will be available.



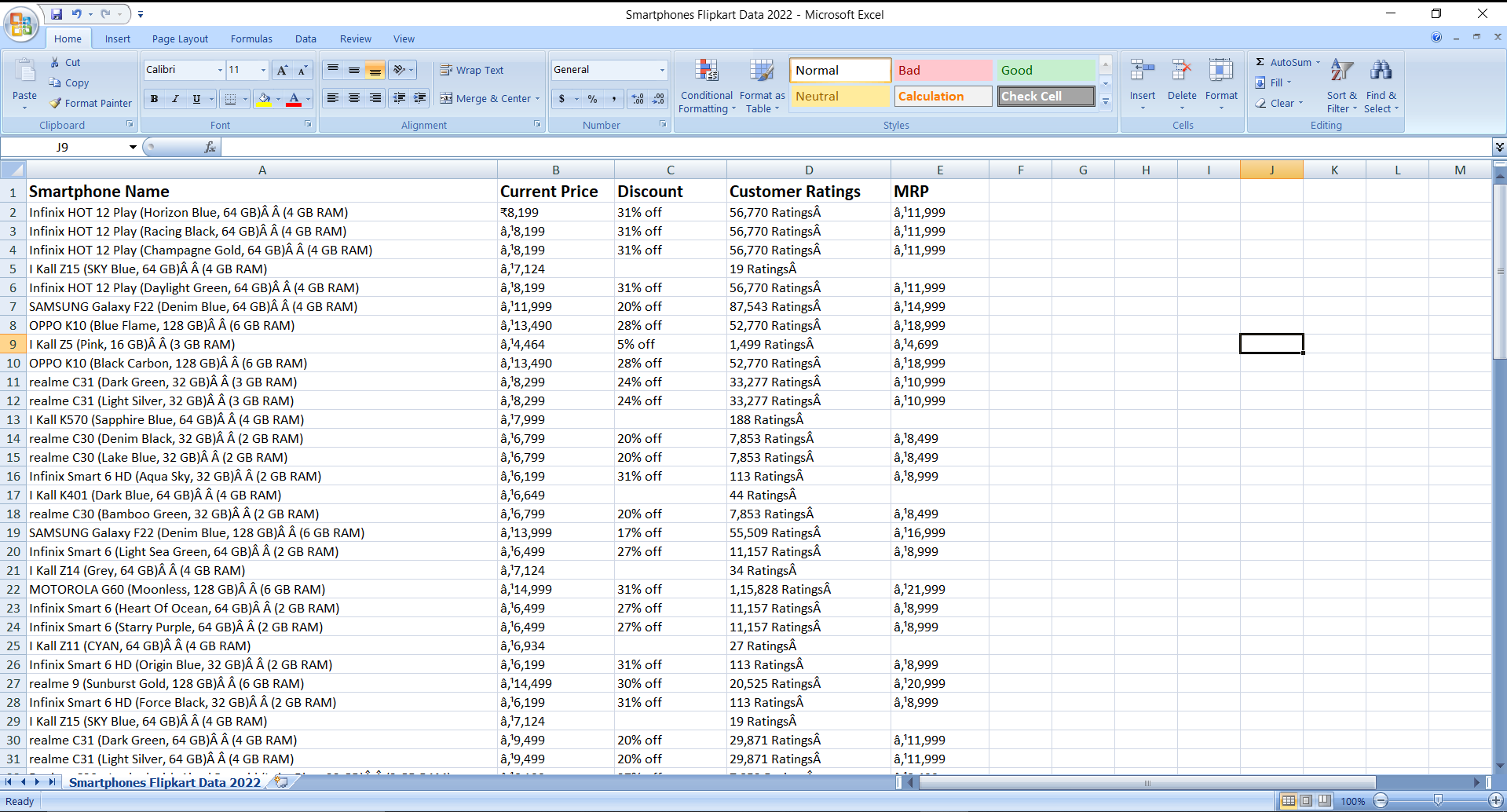
7. Data is being extracted it all depends on us that how much data we want to extract .In this project we are extracting 100 smartphones data so after that press stop .It will show how much data is extracted and how many of them are unique and duplicate. In this we want all data so we click Export all.



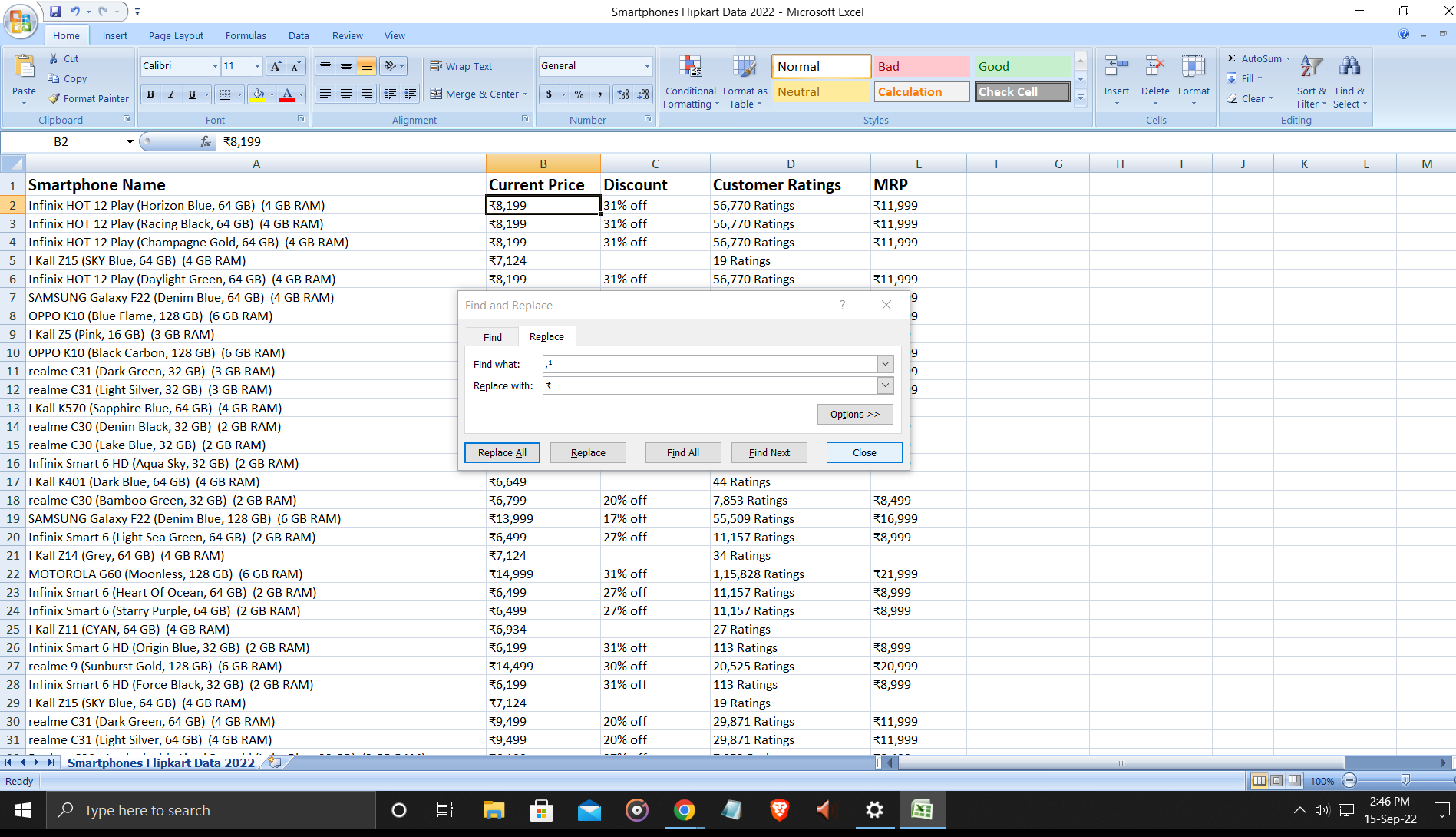
8. It will ask you in which file format you want to store this data we are selecting .CSV you can select based on your preference.



9. Data has been extracted and we can see the data in excel.



10. But there is a little problem in the data ,some data cleaning needs to be done like th pricing is not in INR and some symbol is present in front of that , to remove that we use Ctrl+H in excel and replace the error word with ₹. While we are cleaning the data ,for exact desired output we are exporting data as .xlxs format so that it can be easily readable.



**Our Web scraping is completed and we have the output in rows and columns.**

**SUMMARY**

Finally, we have managed to parse ‘**Flipkart Smartphones Data’** to get our hands on very **interesting and insightful data** when it comes to the smartphone shoping. We have saved all the information we could extract from that website for our needs in a CSV or .xlxs file using which we can further get answers to a lot of questions we may want to ask, eg- if we want to extract the data of the best highly rated headphones we can extract the data easily and put it into csv or .xlxs format.

**FUTURE WORK**

We can now work forward to explore this data more and more to fetch meaningful information out of it. With all the insights , and further analysis into the data, we can have answers to a lot of questions like –

1.Which smartphone has more positive reviews?

2.Which has more discount on mrp?

3.Which brand has most of the smartphones?

4.Which variant of the smartphone is available?

And the list goes on.

In the future, I would like to work to make this DataSet even richer with more data from other lists created by Flipkart.

I would then like to work on analysing the entire data, to know a lot more about smartphones sales than I currently know.

**References**

**1.** Flipkart website [www.flipkart.com](http://www.flipkart.com)

**2**. Octoparse website [www.octoparse.com](http://www.octoparse.com)

**3**. Krish Naik Youtube channel <https://www.youtube.com/watch?v=1PLta8E_HsA&t=35s>

**4**. Web Scraping articles [www.google.com](http://www.google.com)