

**School of Business and Management**

**CHRIST (Deemed to be University), Bangalore**

**MBA 441B**

**Data Exploration and**

**Application using Python**

**END TERM PROJECT**

**TITLE- Web scraping and Exploratory data Analysis of the data**

Submitted to:

**Dr Tripti Mahara**

Submitted by:

**GORRIPATI. LOK SAI**

**2028031**

**Introduction:**

Web scraping is a phrase that refers to a variety of techniques for gathering data from the internet. This is usually done with software that replicates human Web surfing in order to capture certain bits of data from various websites. Web scraping tools may be used by those who want to harvest data to sell to other people or to utilize for promotional reasons on a website. Web scraping, also known as web data extraction, screen scraping, or web harvesting, is a technique for extracting information from the internet.

In today's competitive environment, everyone is looking for new methods to create and capitalise on emerging technology. Web scraping is a technology that allows users to automatically collect structured web data. Web scraping is an excellent choice if the public website you want to get data from doesn't have an API, or if it has but just provides you limited access to the data.

Online scraping is an automated way of extracting structured web data. Web data extraction is another name for it. Pricing monitoring, price intelligence, news monitoring, lead creation, and market research are just a few of the uses for web scraping.

Web data extraction is used by individuals and businesses who want to make better decisions by leveraging the vast amount of publicly available web data.

If you've ever copied and pasted information from a webpage, you've performed the same function as a web scraper, although on a smaller, manual scale. Unlike the time-consuming, mind-numbing process of manually collecting data, web scraping uses smart automation to capture hundreds, millions, or even billions of data points from the internet's seemingly endless expanse.

Data Scraping is a method to recover a lot of information from the web. This method is exceptionally helpful in serious valuing. To actually look at what our item's ideal cost ought to be we can analyze the comparable items that are now on the lookout. These costs can change a great deal. Thus, in this blog, I'm demonstrating how we can scrap information in regard to a specific item.

There are various methods for Data scraping like Scrappy, Beautiful Soup, and so forth the most widely recognized procedure is utilizing BeautifulSoup. It removes the Html page on any site. The information put away by web Scraping is an unstructured organization. We convert the unstructured into organized information for Data investigation and investigation.

The libraries which I have used in this web scrapping and for EDA is shown below:

Requests and Beautiful Soup in Libraries

The Requests module in Python is used to make HTTP requests. It's an easy-to-use library with a lot of features. It will be used to extract material from the URL of a webpage that we wish to scrape.

Beautiful Soup is a library for parsing HTML documents. We create a new Beautiful Soup object by passing the HTML content and the type of parser we want to use.

Pandas: Pandas is a library for data manipulation and analysis. It's used to extract information and store it in the desired format.

Use matplotlib and seaborn for data visualization.

**Objectives:**

The main objective is to find the optimum price range of the televisions which the customers can buy in the flipkart. And, I have performed some more objectives like plotting the histogram, bar graph and scatterplot for the derived columns.

**Challenges faced during collection of data from flip kart using web scrapping:**

* Selecting the web sites from which data can be scrapping
* Some of the websites found difficult to take class.
* One column is having two to three components.
* Splitting the column into 2 categories for many variables
* Data inconsistency
* Null values because of real data
* Wrong data types
* Removing currency values for original and discounted price
* Manual input
* Removing unwanted symbols from the data.
* Data preprocessing is taken more time because the data is real data and having more features to edit and make the whole data as a good data set.

**2.Methodology**

1. **Installing the libraries:**

Graphical user interface, text

Description automatically generated

This will allow us to scrape the data with requests and Beautiful Soup, save the data in a data frame using the Pandas library, and create EDA and visualizations with matplotlib and seaborn.

1. **Creating Empty Lists:**

Text

Description automatically generated

Here we have created a empty list to store the web scraping information in those empty list if there is no any empty list with the required names the data we have gathered will have no storage to save. So it is necessary to create a empty list in the data frame

# Find the URL that you want to scrape:

# A picture containing graphical user interface Description automatically generated

We'll scrape the Flipkart website to get the brand name, prices, and ratings of televisions for this example. This page's address is

[**https://www.flipkart.com/televisions/pr?sid=ckf%2Cczl&wid=18.productCard.PMU\_V2\_9&page={}**](https://www.flipkart.com/televisions/pr?sid=ckf%2Cczl&wid=18.productCard.PMU_V2_9&page=%7b%7d)

# Inspecting the Page: The data is usually nested in tags. So, we inspect the page to see, under which tag the data we want to scrape is nested. To inspect the page, just right click on the element and click on “Inspect”. When we click on the “Inspect” tab, we will see a “Browser Inspector Box” open.

# Graphical user interface, text, application Description automatically generated

# Find the data you want to extract:

# Let's look at the Name, Price, and Rating tags, which are all nested within the "div" tag. We frequently become perplexed as to which tag, we need to copy for this laptop; however, we may use https://webformatter.com/ to filter the tag for Name, price, and rating and copy simply using this application.

# It's time to extract the data from the website now that we've developed the code to access the URL. The data we wish to extract is nested in <div> tags, as previously stated. As a result, I'll look for div tags with those class names, extract the data, and save it in a variable. Refer to the following code:

# Text Description automatically generated with low confidence6. Run the code and extract the data:

# Now the code which is ready to extract the data from the website is ready and also, we have taken only certain classes to work on by running the code we get our required data like the brand name which is pasted its relevant data in their respective classes will store that information in that file name as a column.

A picture containing chart

Description automatically generated

# Store the data in a required format:

# Here the data which is scrapped from the flipkart is stored as a excel csv file and will be stored in the location we have given in the above code

# Preprocessing the scrapped data:

# Now let us check the null values and if there are any null values we can drop those null values because we cant replace values in real data as each product has its own specifications. The data we have gathered is having some coulmns which are not in correct order, so replace those errors in correct way it took a lots of efforts and we have used different codes like replacing and splitting.

# First let us see about the splitting which means that there is two types of data in the same columns we can use that two types of data into two derived columns only but using the split function and I have created the product name, size and display columns using the brand name column. In the same way I have splitted again the rating column into two derives columns like votes and reviews.

# Now let us see the replacing function and how it works this function is used when there is any uwanted symbol or data in the discounted price, original price, exchange offer and exchange price all these are having unwanted symbols and data all these unwanted symbols and data are replaces with just null space.

# Graphical user interface, text, application, email Description automatically generated

# Here we can see in above code that I have called the excel sheet which is already stored in the dataframe. I have got some null values and all those null values are omitted which means all those null values have been removed from the dataset because we cant use those null values even though we can’t replace those null values by mean or median because the each product in flipkart has its own specifications and prices, so we cannot replace those null values other than removing or omitting.

# Below code specifies the replacing and splitting of the columns:

Text, letter

Description automatically generated

Data cleaning and processing is done in above code we have replaced the unwanted symbols with null space and some columns we have split the data into two columns.

First let us see about the splitting which means that there is two types of data in the same columns we can use that two types of data into two derived columns only but using the split function and I have created the product name, size and display columns using the brand name column. Similarly, I've split the rating column into two derivation columns, such as votes and reviews.

Now let's look at the replacing function and how it works. This function is used when there is any unwanted symbol or data in the discounted price, original price, exchange offer, or exchange price. These unwanted symbols and data are all replaced with null space.

A picture containing logo

Description automatically generated

Deleted unwanted columns from the dataframe because we have already formed some new columns from these deleted columns using splitting function.

We have no use of s.no number columns so I have deleted that column and where comes to the brand name I have split that column into three new columns because this brand name has three components in it so I have split that brand name into product name, size and display.



While getting the web scrapping data the order of those columns is in different like the rating and reviews has been in first position and name, pricing is in last position so, I have replaced those order into correct way.

The above code shows the replacing the columns in the data frame in correct order.

Text

Description automatically generated

Later we have checked the data type for each column there are many errors in their data type, and all these columns are having only one data type name as objects and it should be different for different columns based on the data stored in those columns the data type should be there like the numbers should be in integer, decimal values like rating should be in float and factors should be in objects. So, all these error columns having datatype has been replaced with correct data type using above code.

**3.Analysis:**

Pandas describe () is used to view some basic statistical details of a data frame or a series of numeric values, such as percentile, mean, and standard deviation. This method returns a different output when applied to a series of strings, as shown in the examples below. Statistical summary of the data frame as a return type.

Table

Description automatically generated

corr() returns the pairwise correlation of all columns in a dataframe. Any na values are immediately removed from the equation. It is disregarded for any non-numeric data type columns in the dataframe. A variable has a one-to-one correlation with itself.

In this correlation we have taken only numerical columns and made correlation between them, but we have to add all other columns like object datatype as well. For that first we have to convert those object data type with some numerical by using some code.

A picture containing chart

Description automatically generated

Apply () is a function that is used to apply changes to a In Python, this function is equivalent to the map() function. It takes a function as an input and applies it to a Dataframe as a whole. apply function allows users to pass a function and have it applied to each value in the Pandas series. This is a significant enhancement for the pandas library since it allows data to be separated according to the requirements necessary, which is useful in data science and machine learning.



Now let us see the correlation between all the variables in or dataframe including the object data type:

From the output of this correlation, we can say that there is high correlation between votes and review which means that there is more relation between those two variables.

The correlation is between votes and reviews is about 0.885783 and there is another good

correlation between votes and original price and it is about 0.788159.

Chart

Description automatically generated

The below code shows the descending order of the correlation between the variables:

Graphical user interface, text, application

Description automatically generated

Below we can see some graphs which is used for the analysis for getting some information from those graphs like scatterplot, histogram, and bar plot.

The below bar plot is plotted for the original price, here we can say that 20k and 40k is the prices of the televisions that are having highest number of products in the flipkart, and the lowest number of TV’s is present in the prices of 450k, 500k.

Chart, histogram

Description automatically generated

The below histogram shows the ratings of the products and here we can see which rating is given more number of times. The output shows that 4.3 rating is given more times by the customers and the least rating which is given less number of times is 4.9 rating.

Chart, histogram

Description automatically generated

The below histogram shows the votes of the products and here we can see which votes is given more number of times. The output shows that 10k votes is given more times by the customers for a single product and the least votes which is given less number of times is 600k votes for a single product.

Chart

Description automatically generated with medium confidence

This below graph shows the scatterplot between the brand name and the rating of the product. Here we can see that there many brand names and their respective rating we can assume that 4.3 ratings is given commonly for many brands and the least rating is given for the one brand name and the rating is about 2.3.

Chart, scatter chart

Description automatically generated

The below bar plot is between rating and the votes. Here we can see below output graph shows that rating of 4.4 is having high number of votes which is above 40k and the least rating of 4.1 is having with the votes of 1000.

Chart, histogram

Description automatically generated

The below bar plot is between discounted price and size. Here we can see below output graph shows that 43inch size of the tv is having high number of discounted prices which is above 250k and the least tv size of 53inch size is having of 2k discounted price.

Chart, bar chart

Description automatically generated

The below bar plot is between discounted price and display. Here we can see below output graph shows that 4k Led tv is having high number of discounted prices which is above 200k and the least smart tv display of smart Led is having of 3k discounted price.

Chart

Description automatically generated

The below bar plot is between exchange price and percentage offer. Here we can see below output graph shows that 47 percent offer is having highest number of xchange prices which is above 12k and the least 8 percent offer is having of 1.5k exchange price.

Chart, bar chart

Description automatically generated

The below graph shows the bar plot between the discounted price and the rating, now we can see that five rating is given only for the products 1,55,000. And the least rating 2.3 is given for the discounted price of 3,40,000.

Chart, bar chart

Description automatically generated

Now the last analysis which is important for our objective which is used to find the optimum price where the customers can purchase the tv in the flipkart. For this we have implemented the boxplot from that we got the outliers.

As can be seen, there is one outlier with an extremely wide price range.

We came to the following conclusions: Based on the above visualization, my analysis is that the product is available in the 2.5,2.8,3.9,4.0,4.1,4.2,4.3,4.4,4.5 rating range with a price range of 20,000 to 60,000 rupees.

Chart

Description automatically generated

**4.Recommendations:**

There are few recommendations I would like to give for the flipkart based on the product quality, packing of the Tv product and performance.

Product quality:

There are some questions asked by the customers to the flipkart based on their product quality like Is the material of the product good and long-lasting? Is the vendor reputable? What has been the response from other people who have purchased it? These are some of the quality-related questions that customers frequently ask themselves before purchasing a product. They should consider these ‘Quality' signals and place higher-quality goods at the top of the Recommended list. This contributes to the consumers' ‘trust factor' when it comes to the suggestions.

Product Performance:

The more popular and quickly sold a product is, the greater the user's odds of making a purchase choice are. To guarantee that goods with high performance are presented to the consumer, we check previous performance and popularity as one of the signals in the ranking. It will help to purchase that product.

Packing of the product:

When the product is delivered to the consumer, the appearance and feel has a significant influence on the buyer's experience. The use of proper packaging material while shipping items reduces the risk of damage while in transport. More importantly, it shows great service and builds customer trust in you and your company, which may lead to more sales.

**5.conclusion:**

The conclusion for this web scrapping is that we collected the data from the flipkart website and made required number of derived columns which was difficult to gather all that data and there is more preprocessing and cleaning work on the web scrapping data.

The product I have chosen is televisions in flipkart and with the data collected I have performed some EDA on the data using Bar plot, histogram, scatter plot and boxplot.

I have performed the Boxplot to identify which price range is more suitable for the purchasing and I got many outliers which are above 100,000 rupees by this we can say that the televisions whose price ranges between 20,000-60,000 are the best for purchasing.