**DECLARATION**

I, hereby, declare that the work presented in this study or thesis report is the outcome of the design and development work performed by me under the supervision of **Muhammad Shafayat Oshman**, Associate Professor, Electrical & Computer Engineering, North South University as a course work of CSE/EEE 498R (Direct Research). We also declare that no part of this report has been taken from other works without reference.

Signature of Students

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Md.Muttashim Mishel Mawn**

**APPROVAL**

This Thesis report titled ‘Data Scraping for Competitive Market Analysis and Behavior’ submitted by Md.Muttashim Mishel Mawn ID: 1520800042 to the Department of Electrical and Computer Engineering, North South University, has been accepted as Direct Research Term Final Report.

Signatures

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Muhammad Shafayat Oshman**

Associate Professor, ECE

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Dr. Rajesh Palit**

Associate Professor and Chairman, ECE

**Abstract**

I have proposed a data analysis based study for a competitive market analysis from our country’s online based market. Because of now a days a lot of online sale service provides various product at our door but it’s so difficult for a newcomer to establish a sustaining business. Competitive analysis is one of the many aspects of digital marketing. It also requires data from various dynamic website. The main objective of our study is to develop a crawler to gather data to compare them and bring up a result that shows us the market status, which type of product they stock most and the demand of a specific product. And for this project I am going to use python library bs4 and scrappy framework for data scraping after that for data filtering I am going to use python pandas. The outcome of this project will help the new comers in the market to understand market status and consumer also can find out the availability of a certain product and what are their services according to specific brands and products.

**Table of Content:**

|  |  |
| --- | --- |
| **Title** | **Page No** |
| Abstract | 3 |
| Table of Content | 4 |
| Introduction | 5 |
| Problem Statement | 5 |
| Description | 6 |
| Methodology | 6 |
| Technology | 11 |
| Discussion | 11 |
| Future Work | 11 |
| Cost | 12 |
| Conclusion | 12 |

**Introduction:**

In 2022 business is more competitive than 10 years before. Advanced technology gives us the path and the solvent. In Bangladesh business for a newcomer is way harder than the traditional family business man who took the business experience from his father and his father took the business experience from his grandfather. And this is one of the reasons that our new generation is scared to start a new business.

Right now the market is competitive. No one gives you the secret sauce to get success. And I can say competitive market analysis is one of the most important aspects of digital marketing. Maybe everyone can get the initial success by following basic steps of starting a business but what about sustaining business development. If we talk about sustaining business development we should focus on product analysis and need to read consumer philosophy and their capability.

**Problem Statement:**

Nowadays the competitive market is not newcomer friendly. A lot of newcomers just started their first business but that cannot even sustain. The reason behind this is not enough knowledge about market status, product analysis and consumer psychology to buy a product from you. My approach is to develop a crawler that will scrape data from various ecommerce websites and compare products by price range, demand and supply ratio, services that sellers provide us and how often consumer’s write a review. If we are able to compare this we can get the data of the number of stock products in a specific price range. So we can determine the demand of a specific product. After all, it will help newcomers and even the experienced business man too.

**Description:**

A competitive analysis is a strategy that involves researching major competitors to gain insight into their products, sales, and marketing tactics. Implementing stronger business strategies, warding off competitors, and capturing market share are just a few benefits of conducting a competitive market analysis.

Data scraping involves pulling information out of a website and into a spreadsheet. To a dedicated data scraper, the method is an efficient way to grab a great deal of information for analysis, processing, or presentation.

People can analyze data by themselves but it's not efficient to do manually in 2022. So what if a bot does the whole work and gives you the outcome and does not consume more time? So I want to create a crawler that crawl data from various websites and gathers data in a single csv file. Here you can gather data as much as you can. More data will give you a more appropriate outcome in any situation. The main difference between an experienced businessman and newcomer is the experienced businessman has better knowledge about product and consumer. Even newcomers know those facts but data changes over time so everyone should change their strategies over time too. And this is the main weak point for newcomers. If you fail to adopt you can't survive long. For example: when the rainy season has come the number of umbrella sales increases drastically everyone knows it but it does not depend on the rainy season it depends on how often rain comes. So if you stock a pile of umbrella at the beginning of the rainy season, then it can be said that you have adopted such a strategy due to lack of data analysis of your previous year.

Data and statistics will predict the future. So in this competitive market if you want to survive there is no other way then market and product analysis.

**Methodology:**

My approach to solving this problem is to break the problem down into smaller parts and solve them one by one. The first thing that I need is information. So I decided to scrape data from the top five computer hardware and peripheral shops in Bangladesh and select five types of products from their website. Collecting data from various websites is not that much hard but when I crawled data from websites it will take so much unnecessary stuff and it will prevent our next step.



Fig.1 Corrupted Data

So I need to organize all of the data in a uniform format. The more we clean the data the more we will get the accurate outcome. So that data cleaning is the most important section for any kind of data analysis project. After that we need to remove all of the corrupted data from the data list. For this project we are storing product name, product price, product category, product link and product review. If any one of the variables is missing we can call it corrupted data. Because when we will start to filter and compare one with another our project outcome accuracy will decrease. But i didn’t edit our main dataset; we just created a fresh new dataset in another file. Finally we have a dataset to work with.

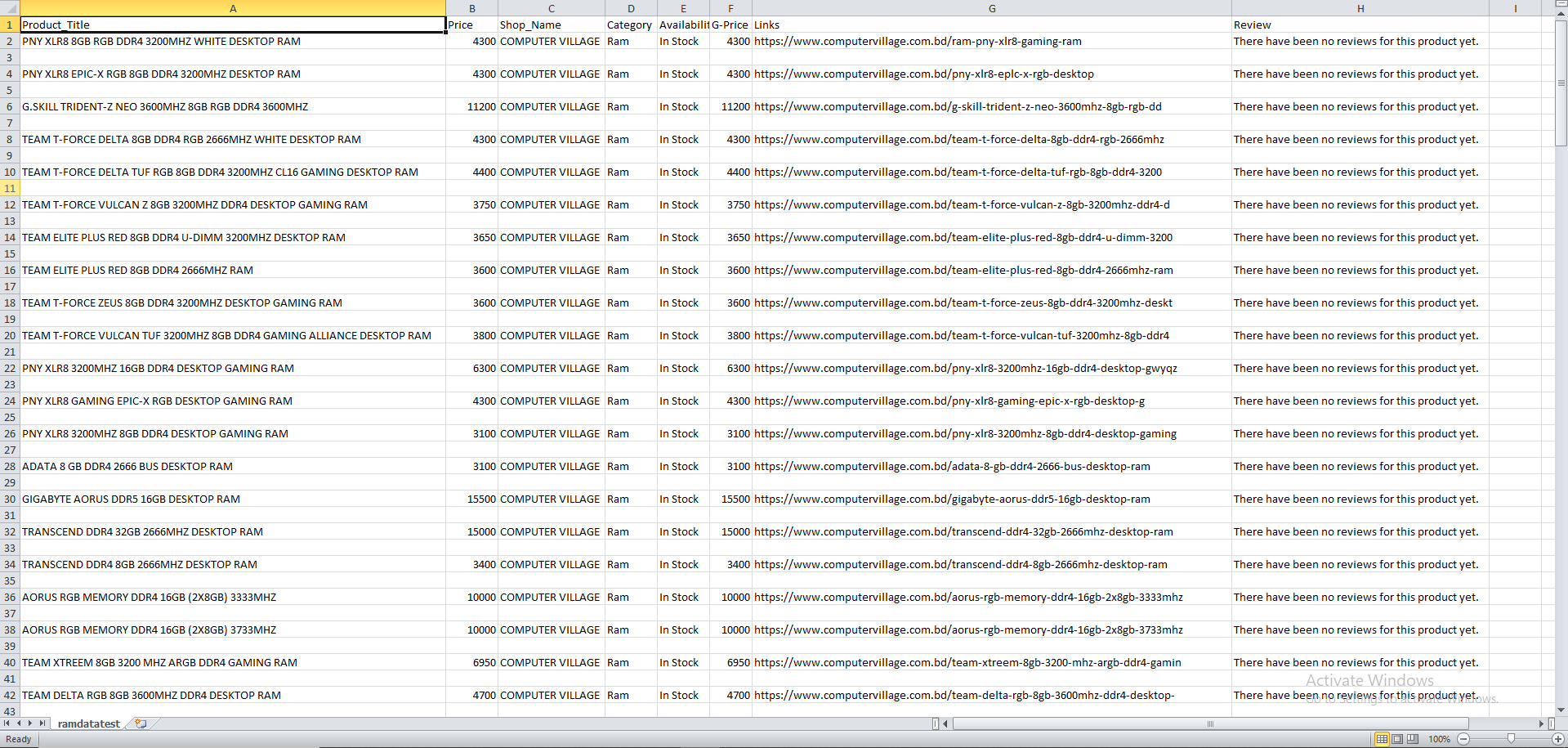


Fig. 2 Dataset

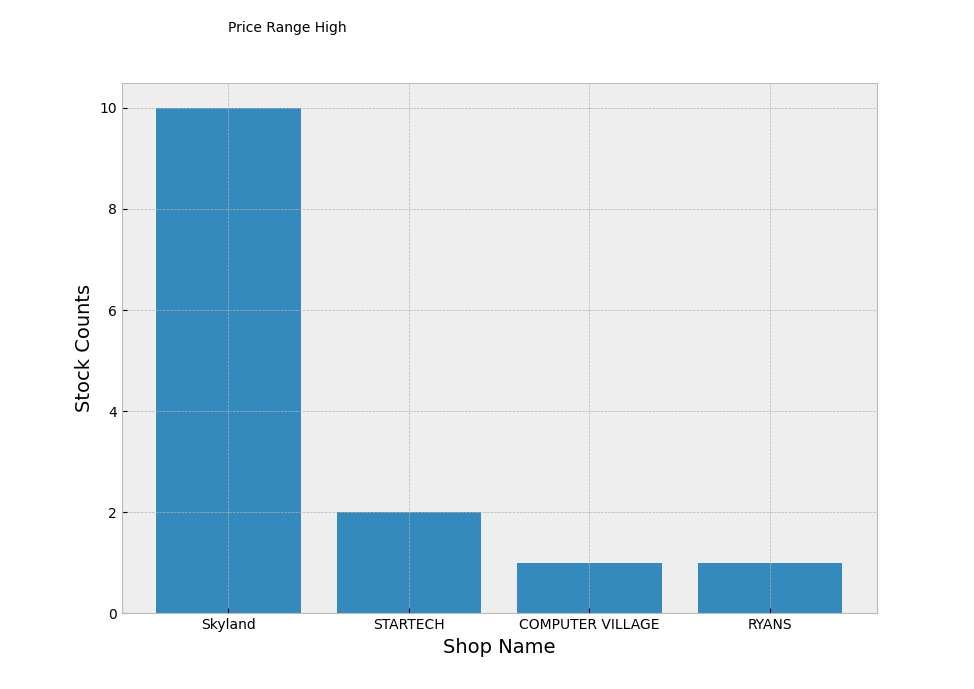
Now we are in data analysis. First we will see how the product stocks within a certain price range. And to determine the range of this product, I extracted the minimum price and maximum price of the product from my data set and divided the price between them into two ways, one lower price range and the other upper price range.   
  


Fig.2 Stock counts in high price range

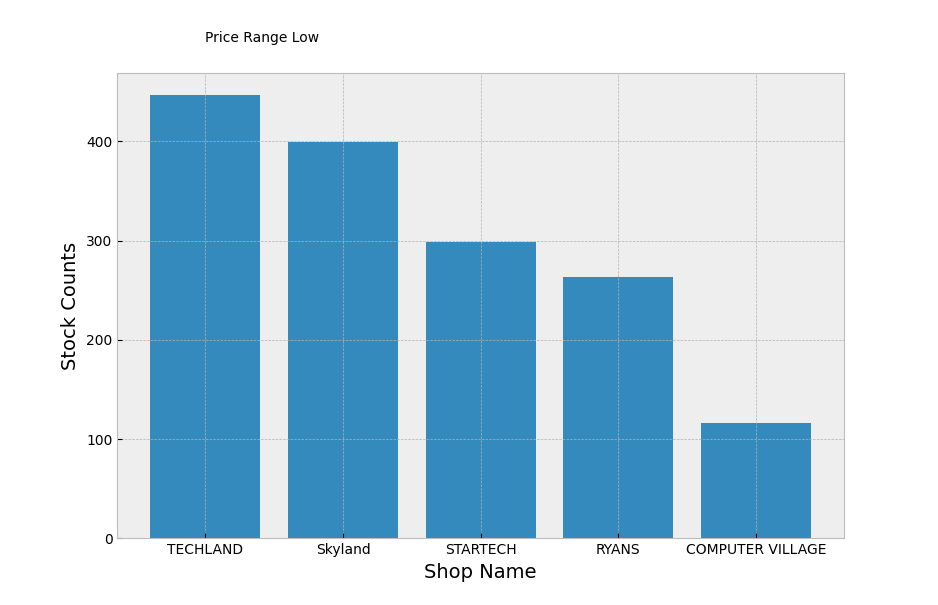


Fig.3 Stock count in low price range

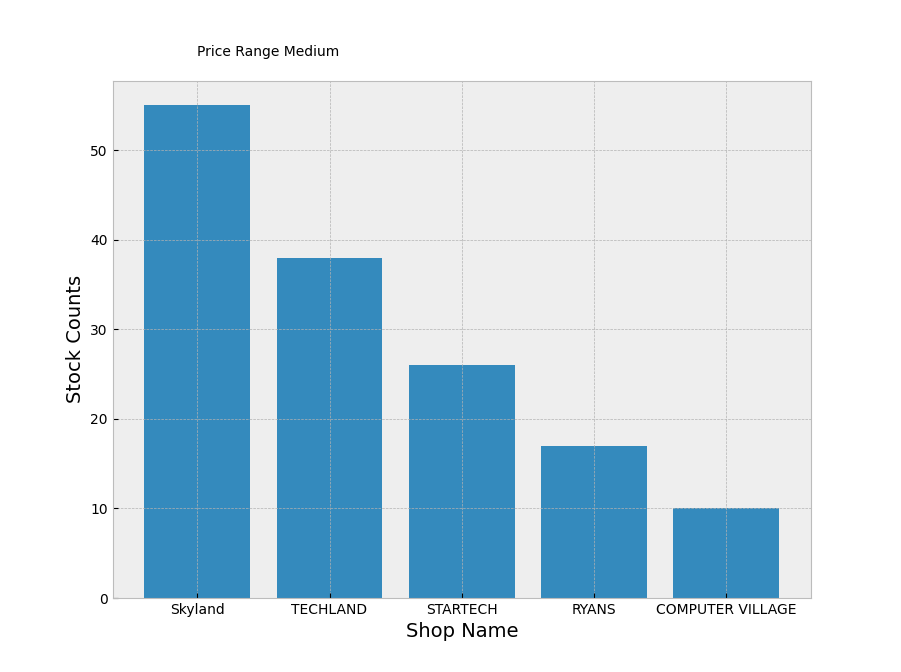


Fig.4 Stock Counts in medium price range

We can understand from the three bar charts that the stocks in the lower price range products are more than the higher price range products. And it is so obvious for any type of market. So we need a specific and more meaningful outcome. And for that i extracted the minimum price range, medium price range and high price range from the data set. And this time I found different results from different product categories. And the outcome will change again over time. Maybe in the future the number of medium price ranges will increase with demand. Because In economics. An increase in supply, all other things unchanged, will cause the equilibrium price to fall, quantity demanded will increase. A decrease in supply will cause the equilibrium price to rise, quantity demanded will decrease.

We have one more variable that will provide us with consumer psychology or trends of any kind of product. If we compare the number of reviews and the number of products within a certain product price range, we will know the consumer's soft spot and which products are more demanded in a specific time period.

**Technology:**

For this problem I am using python programming language as my base technology. But I am using different libraries for the rest of the work. They are -

1. Beautiful soup 4
2. Pandas
3. Numpy
4. Matplotlib

**Discussion:**

Initially I tested 5k data, measured the outcomes and then I tested 10k data and measured the outcomes. And what I saw the result changed when the number of data increased. One more thing I tried during this time period is scraping the same data set in two different time periods but I cannot find any changes because the owner of those websites didn’t update their stock and upcoming product information. I can assure that if i can scrape data from two different time periods we will get a more appropriate result.

**Future Work:**

Basically we have data so we can solve different types of problems too. If we compare individual products in different shops we can find out who provided the minimum price of a specific product. It will help the consumer because comparing a single product from different websites is more time consuming. Also everyone knows that some bad businessmen create a syndicate and they increase a certain product's demand and price by holding products. So for this problem I need both price data that is specified by the company and the vendor. It will help to break the unnecessary price hike.

**Cost:**

For this project we just need a better computer to scrape more data in a short time. I scraped 16k data and my computer takes half an hour because I have a mid-end computer. But if we will scrape over 1m data we definitely need high end devices because when i start to scrape multiple pages from a single website  i need to put a delay after every page so that server cannot detect it as a bot. If the server detects it, the server will block it. So if we want to avoid this issue we just need a high end device. After that it won’t take too much time to execute.

I am working on a computer that costs around 2lakhs. If we want to run this project for industrial purposes we need a computer that is around 5 -7 lakhs. Because we need a faster processing unit.

**Conclusion:**

In this study I am trying to gather some information about the current market situation that will influence our newcomers to become entrepreneurs. It will not only help the individuals but also puts a huge impact on our central economy, unemployment and a better environment for consumers to buy products.