Tunisian Republic
Ministry University of Higher Education and Scientific Research
Carthage University – Engineering School of Statistics and Information Analysis

End Of Year Project Report

The Height School of Statistics and Information Analysis

Submitted By

Nour Sfar

Web Scrapping and Price Benchmarking

February – May 2023

Supervisor:

Mrs. Haifa Ben Massoud

Thanks

First of all, I would like to thank my project supervisor, Mrs. Haifa BEN MASSOUD, for her guidance throughout the work period, as well as for taking the time to supervise the progress of my end-of-year project. Finally, I would also like to thank all the people who participated in the realization of this work.

Abstract

This work is part of the end-of-year project carried out at the Height School of Statistics and Information Analysis. The objective of this project was to carry out a comparative analysis of the different competing brands of a product of our choice. The project was carried out in two main stages: the first stage consisted of collecting the necessary data for each product using the **“web scraping”** technique. The second step was to visualize this data on a **“Dashboard”** that we created to analyze it and carry out our comparative study.

Résumé

Ce travail s'inscrit dans le cadre du projet de fin d'année réalisé au sein de l'Ecole Supérieure de la Statistique et de l'Analyse de l'Information. L'objectif de ce projet était de réaliser une analyse comparative des différentes marques concurrentes d'un produit de notre choix. Ce dernier s'est déroulé en deux grandes étapes : la première a consisté à collecter les données nécessaires pour chaque produit en utilisant la technique du **“web scraping”**. La deuxième étape a été de visualiser ces données sur un **“tableau de bord”** que nous avons créé pour les analyser et réaliser notre étude comparative.

Table Of Content

Introduction

As part of my studies at the School of Statistics and Information Analysis, I was asked to do a final year project. A **comparative analysis** of headphones prices is the subject of the project. Comparing competitive prices between different major brands, with the aim of understanding the competitive position of one's own offering in relation to that of competitors. A technique that allows the company to conquer a market as powerful and dynamic as the headphones market.

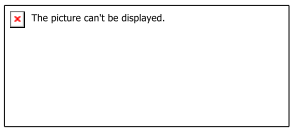
As part of this project, I learned about **web scraping** and its basic tools and how to use it to perform a comparative price analysis. In addition, this internship is an opportunity to strengthen my practical skills as an engineer in **statistics and information analysis**, moreover, I had the advantage of practicing computer languages, especially python, as well as practicing the **descriptive statistics** modules learned during the first year of engineering studies. This document contains a presentation of different brands headphones.

Chapter 1

Presentation of The Research Product

1.1 Presentation of the “MyTek” Web-site:

Mytek.tn is an online shopping website based in Tunisia that offers a wide range of products, including electronics, appliances, computers, smartphones, and more. The website features a user-friendly interface that allows customers to browse and search for products easily. It offers multiple payment options, such as online payment, cash on delivery, and bank transfer. Mytek.tn also provides various delivery options, including home delivery and in-store pickup. The website is known for its competitive prices, regular promotions, and excellent customer service. Additionally, it has a dedicated customer support team that is available to assist customers with their inquiries and provide after-sales support. Overall, Mytek.tn is a reliable and convenient online shopping destination for Tunisian consumers.



1.2 Product Description:

Earbuds, are a type of small and lightweight headphones that fit directly into the ear canal. They are designed to provide high-quality audio and a comfortable fit, making them ideal for use during exercise, travel, or everyday activities. Earbuds come in various shapes and sizes. Wireless earbuds are becoming increasingly popular due to their convenience and ease of use, with many models featuring Bluetooth connectivity and advanced features like noise-cancellation, touch controls, and voice assistants. Earbuds are compatible with a wide range of devices, including smartphones, tablets, laptops, and music players, making them a versatile and essential accessory for anyone who enjoys listening to music or podcasts on-the-go.

1.3 Brands Competitors:

The market for Bluetooth headphones has become increasingly crowded in recent years, with several big brands competing for market share. Apple's AirPods are arguably the most well-known Bluetooth earbuds, offering seamless integration with Apple devices, long battery life, and a compact design. Samsung's Galaxy Buds and Sony's WF-1000XM4 are two other popular options, both featuring noise-cancellation, long battery life, and impressive sound quality. Other major players in the Bluetooth headphone market include Bose, Jabra, and Sennheiser, all of which offer a range of earbud and headphone models with different features and price points. Additionally, there are several budget-friendly options available from brands like Anker and Sound core, offering reliable performance at a lower cost. With so many options available, consumers have a wide range of choices when it comes to selecting Bluetooth headphones that best suit their needs and preferences.

Chapter 2

Web Scraping

2.1 Introduction:

Web scraping involves extracting data from websites and saving it for analysis or other uses. Scraping allows for the collection of various types of information, such as contact information like email addresses or phone numbers, as well as individual keywords or URLs. This information is then gathered into local databases or data frames, as was done in this study. This chapter serves as preparation for the study, where I present the different libraries used, methods, and approach for extracting and cleaning the data.

In this study, I used the Python language, intended for statisticians and data scientists, as well as five libraries, including libraries for extracting and cleaning the database.

2.2 Libraries:

The first library, Beautiful Soup, is a library that serves as a browser automation tool. It provides extensions to replicate user interactions with browsers and is exclusively based on HTML and JavaScript. The second library, rvest, allows for parsing the content of a web page to make it usable by R. For example, it is possible to create a list from a Wikipedia page, retrieve text from a page, or transform the extracted HTML table into a data frame.

Beautiful Soup is a Python library for parsing HTML and XML documents created by Leonard Richardson. It produces a syntax tree that can be used to search for or modify elements. When the HTML or XML document is poorly formed (for example, if it lacks closing tags), Beautiful Soup uses a heuristic-based approach to reconstruct the syntax tree without generating errors. This approach is also used by modern web browsers. These libraries were installed in R through the “install.package(...)” command. The libraries used allowed for web scraping and extraction of the necessary data for the study. They were also used to clean the data before analyzing it.

2.3 Phases:

2.3.1 Data Extraction:

The simplest way to install not only BeautifulSoup but also Python and its most popular packages such as Python, NumPy, and Matplotlib, is to use Anaconda, a multi-platform (Linux, macOS, Windows) Python distribution for data analysis and scientific computing.

To begin, we will open a Jupyter notebook and import the libraries that we have installed (except for lxml, which does not need to be imported). Then, we are ready to retrieve our first web page. It's not very complicated: we save the URL we want to scrape in a variable called "URL", and then we send a request to retrieve the HTML content of the web page with "requests.get(URL)".

If you print the "response" variable, you will see that the HTTP response status code is 200, which means that the URL request was successful. However, we need the HTML content of the requested web page. The next step is to extract the HTML content from "response" and save it in a variable called "html".

The result is the HTML content of the MyTek.tn list pages, but it is really hard to read for the human eye. Luckily for us, we have Beautiful Soup and lxml!

2.3.2 Data Cleaning:

Extracting data can be complex; however, it is not the most challenging step because the format of the extracted information is not always optimal. The data may contain unnecessary characters that can lead to errors during statistical analyses. Therefore, it is necessary to remove outliers, eliminate duplicates and irrelevant information from the data.

2.3.3 Final Database:

type	Gamer	Brand	Bluetooth Gen	Spread	Color	Availablity	Price(DT)
EarBuds	no	BELKIN Soundform	Unknown	Unknown	Black	In stock	199
EarBuds	no	XIAOMI	5	10	Pink	In stock	215
EarBuds	no	BEATS	Unknown	Unknown	Gold	In stock	349
EarBuds	no	HUAWEI	5.2	Unknown	Unknown	In stock	299
EarBuds	no	JBL Tune_Flex	5.2	Unknown	Black	In stock	355
EarBuds	no	KSIX	5.0	Unknown	Black	In stock	269
EarBuds	no	KSIX	5.2	Unknown	Gray	In stock	229
EarBuds	no	TWS	5.0	10	Gold	In stock	79
EarBuds	no	OPPO Enco_Air	5.2	Unknown	White	In stock	279
EarBuds	no	BEATS	Unknown	Unknown	Gray	In stock	349
EarBuds	no	OPPO Enco_Buds	5.2	Unknown	White	Incoming	149
EarBuds	no	Fantasy	Unknown	Unknown	White	In stock	119
EarBuds	no	ONEPLUS	5.0	10	White	In stock	349
EarBuds	no	TWS	5.0	10	Rouge	In stock	79
EarBuds	no	ONEPLUS	5.0	10	Black	In stock	159
EarBuds	no	ONEPLUS	5.2	10	White	In stock	499
EarBuds	no	ONEPLUS	5.2	10	Black	In stock	499
EarBuds	no	NOKIA Power_Earbuds	5.0	Unknown	Fjord	In stock	199
EarBuds	no	5_Plus_Pro	Unknown	Unknown	White	Incoming	69
EarBuds	no	XIAOMI Redmi_Buds3	5	10	White	Sold out	159
EarBuds	no	Galaxy Buds_Pro	Unknown	Unknown	Black	Sold out	479
EarBuds	no	Galaxy Buds2	Unknown	Unknown	Olive	Sold out	429
EarBuds	no	JBL Reflect_Mini	Unknown	Unknown	Beige	In stock	445
EarBuds	no	JBL Reflect_Mini	Unknown	Unknown	Black	In stock	445
EarBuds	no	MARSHALL	5.2	10	Black	In stock	489
EarBuds	no	ANKER	5.0	Unknown	Black	In stock	399
EarBuds	no	NOKIA Power_Earbuds	5.0	Unknown	Black	In stock	269
EarBuds	no	CONTACT	5.1	10	Black	In stock	69
EarBuds	no	XIAOMI Redmi_Buds_Essential	5.2	Unknown	Black	In stock	69
EarBuds	no	CONTACT	5.1	10	White	In stock	70
EarBuds	no	XIAOMI MIBRO_Earbuds3	5.3	Unknown	White	In stock	99
EarBuds	no	XIAOMI MIBRO_Earbuds3	5.3	Unknown	Pink	In stock	99
EarBuds	no	Galaxy Buds_Pro	Unknown	Unknown	Purpule	In stock	499
EarBuds	no	HUAWEI	5.2	Unknown	Blue	In stock	219
EarBuds	no	Galaxy Buds2	Unknown	Unknown	Purpule	In stock	499
EarBuds	no	Galaxy Buds_Pro	Unknown	Unknown	Gray	In stock	479
EarBuds	no	OPPO Enco_Buds2	5.2	Unknown	Blue	In stock	99
EarBuds	no	OPPO Enco_Air2	5.2	Unknown	White	Incoming	169
EarBuds	no	OPPO Enco_Air3	5.3	Unknown	Purpule	In stock	199
EarBuds	no	OPPO Enco_Air3	5.3	Unknown	White	In stock	199

Chapter 3

Dashboard

In this chapter, I will begin by providing a brief introduction to the Power BI software and then present the comparative analysis dashboard.

3.1 Power Bi:

Power BI Desktop is a data analysis application developed by Microsoft. The software enables users to connect to data sources, transform and visualize data, and create customized dashboards and interactive reports. These visualizations can be published to a workspace and shared among multiple users. In addition to the existing features such as graphs and maps, the software also provides visual elements for R or Python scripts that can be useful for developers.

To get started, users need to import their working databases into Power BI. The software offers a variety of database file types, including Excel, SQL Server, and datasets. It also allows users to make modifications to the columns in the database.

3.1 The Dashboard:

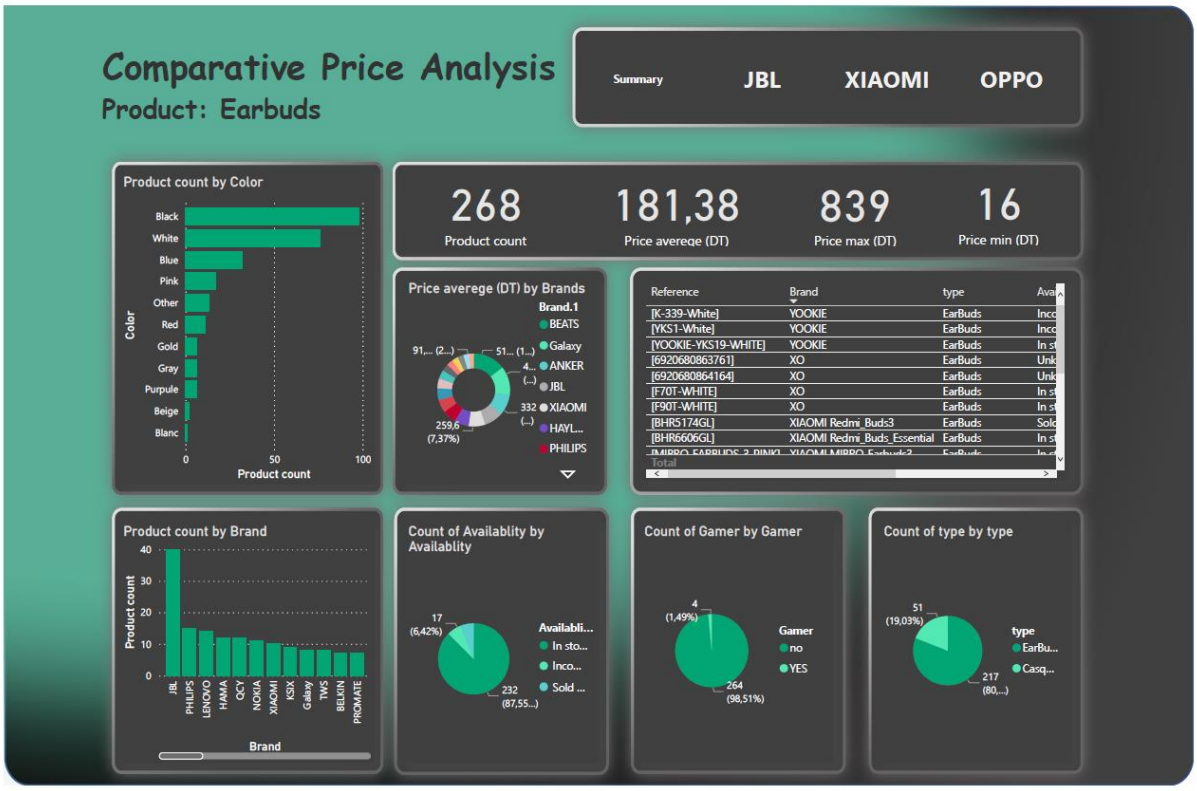
The dashboard created consists of four pages. The first introductory page contains the title and two buttons: The first button leads to a page containing visualizations of the data in the form of graphs, which are linear representations of the different variables in the database.

Conclusion

This end-of-year project has been rewarding both from a practical and personal standpoint. On a practical level, it has allowed me to showcase and enhance my programming skills in the Python language. I also discovered and used for the first time the Power BI dashboarding tool, as well as the various means and appropriate graphics to perform comparative analysis. Additionally, the difficulties encountered in web scraping were useful in broadening my knowledge in this field and discovering the different methods used by top web developers to overcome these obstacles.

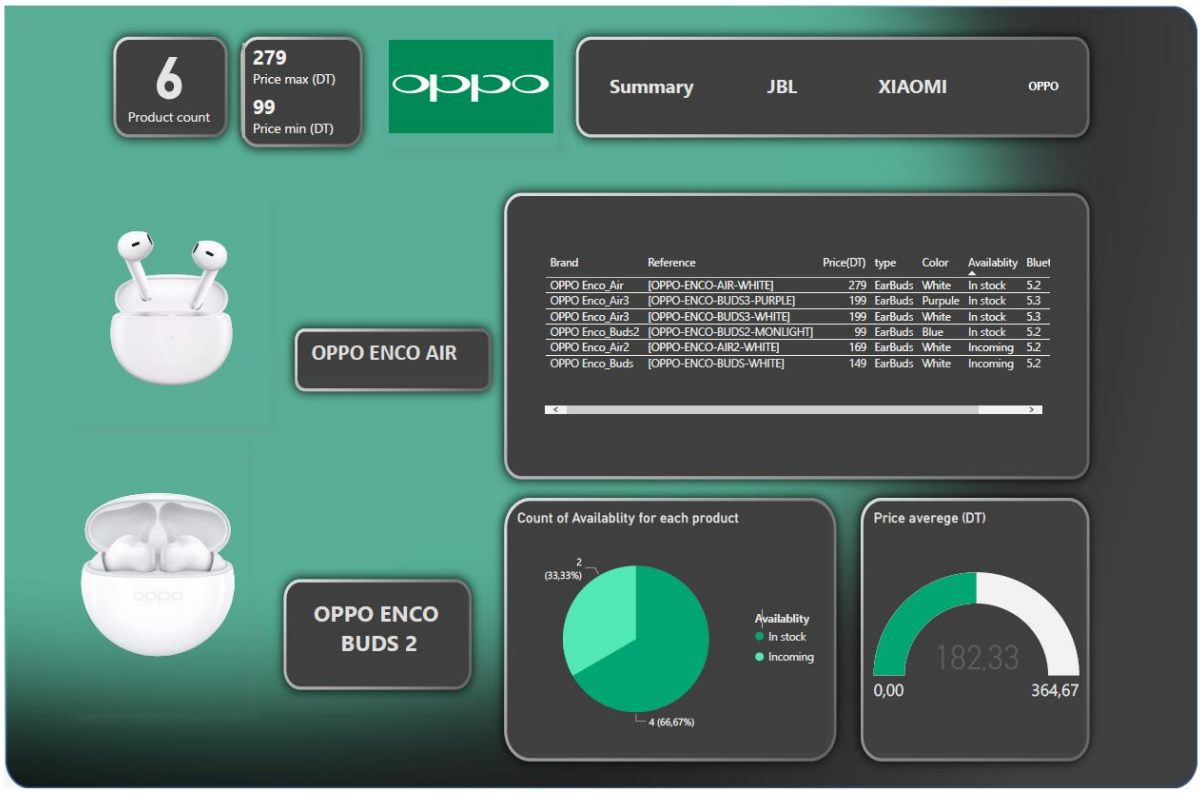
From a personal perspective, this project allowed me to practice oral communication through weekly presentations, as well as develop good time management practices. The work was divided over several weeks of the semester. Finally, it is evident that the work done is only a primary step for more in-depth projects.

Attachment


Comparative Price Analysis
Product: Earbuds
Summary
JBL
XIAOMI
OPPO
Product count by Color
Black
White
Blue
Pink
Other
Red
Gold
Gray
Purpule
Beige
Blanc
Color
0
50
100
Product count
268
Product count
181,38
Price averege (DT)
839
Price max (DT)
16
Price min (DT)
Price averege (DT) by Brands
Brand.1
BEATS
Galaxy
ANKER
JBL
XIAOMI
HAYL...
PHILIPS
91,... (2...)
51... (1...)
4...
(...)
332
(...)
259,6
(7,37%)
Reference
Brand
type
[K-339-White]
YOOKIE
EarBuds
[YKS1-White]
YOOKIE
EarBuds
[YOOKIE-YKS19-WHITE]
YOOKIE
EarBuds
[6920680863761]
XO
EarBuds
[6920680864164]
XO
EarBuds
[F70T-WHITE]
XO
EarBuds
[F90T-WHITE]
XO
EarBuds
[BHR5174GL]
XIAOMI Redmi_Buds3
EarBuds
[BHR6606GL]
XIAOMI Redmi_Buds_Essential
EarBuds
Total
Product count by Brand
40
30
20
10
0
Product count
JBL
PHILIPS
LENOVO
HAMA
QCY
NOKIA
XIAOMI
KSIX
Galaxy
TWS
BELKIN
PROMATE
Brand
Count of Availablity by Availablity
17
(6,42%)
232
(87,55...)
Availabli...
In sto...
Inco...
Sold ...
Count of Gamer by Gamer
4
(1,49%)
264
(98,51%)
Gamer
no
YES
Count of type by type
51
(19,03%)
217
(80,...)
type
EarBu...
Casq...


40
Product count
735
Price max (DT)
125
Price min (DT)
JBL
by HARMAN
Summary
JBL
XIAOMI
OPPO
Brand
Reference
Bluetooth Gen
type
Color
Average of
JBL Quantum_610
[JBLQUANTUM610BLK]
Unknown
Casque Bluetooth
Black
JBL Live_Pro+
[JBLLIVEPROPTWSBLK]
5.0
EarBuds
Black
JBL Live_Pro
[JBLLIVEPROPTWSPIK]
Unknown
EarBuds
Pink
JBL Live_Free
[JBLLIVEFRNCPTWSU]
Unknown
EarBuds
Blue
JBL Live_Free
[JBLLIVEFRNCPTWSW]
Unknown
EarBuds
White
JBL Reflect_Mini
[JBLREFLMININCBLK]
Unknown
EarBuds
Black
JBL Reflect_Mini
[JBLREFLMININCWHT]
Unknown
EarBuds
Beige
JBL Tune_Flex
[JBLTFLEXBLK]
5.2
EarBuds
Black
JBL Live_500BT
[0094036]
Unknown
Casque Bluetooth
White
JBL Live_500BT
[JBLLIVE500BTBLK]
Unknown
Casque Bluetooth
Black
JBL Live_500BT
[JBLLIVE500BTBLU]
Unknown
Casque Bluetooth
Blue
JBL Live_500BT
[JBLLIVE500BTRED]
Unknown
Casque Bluetooth
Rouge
JBL Free_II
[JBLFREEIITWSWHT]
5.0
EarBuds
White
Total
JBL tune T510
JBL Wave 300 TWS
Count of product by type
12
(30%)
28
(70%)
type
EarBuds
Casqu...
Count of Gamer Products
1 (2,5%)
39
(97,5%)
Gamer
no
YES
Price averege (DT)
286,20
0,00
572,40


10
Product count
549
Price max (DT)
59
Price min (DT)
Summary
JBL
XIAOMI
OPPO
XIAOMI MI Earbuds Basic
XIAOMI buds 3t pro
Brand Price(DT) Reference Availablity Bluetooth Gen Color Spread type
XIAOMI 215 [BHR4798CN] In stock 5 Pink 10 EarBu
XIAOMI Buds_3T_Pro 549 [BHR5177GL] In stock 5.2 White 10 EarBu
XIAOMI Buds_3T_Pro 549 [BHR5275GL] In stock 5.2 Black 10 EarBu
XIAOMI Buds_3T_Pro 549 [BHR5917GL] In stock 5.2 Green 10 EarBu
XIAOMI HAYLOU_T16 249 [HAYLOU-T16] In stock 5 Black 10 EarBu
XIAOMI Mi_Earbuds_Basic 59 [BHR4272GL] In stock 5.0 Black 10 EarBu
XIAOMI MIBRO_Earbuds3 99 [MIBRO-EARBUDS-3-PINK] In stock 5.3 Pink Unknown EarBu
XIAOMI MIBRO_Earbuds3 99 [MIBRO-EARBUDS-3-WH] In stock 5.3 White Unknown EarBu
XIAOMI Redmi_Buds_Essential 69 [BHR6606GL] In stock 5.2 Black Unknown EarBu
XIAOMI Redmi_Buds3 159 [BHR5174GL] Sold out 5 White 10 EarBu
Count of of product by Bluetooth Gen
1 (10%)
2 (20%)
4 (40%)
3 (30%)
Bluetooth Gen
5.2
5
5.3
5.0
Price averege (DT)
259,60
0,00
519,20


6
Product count
279
Price max (DT)
99
Price min (DT)
Summary
JBL
XIAOMI
OPPO
OPPO ENCO AIR
OPPO ENCO BUDS 2
Brand Reference Price(DT) type Color Availablity Bluet
OPPO Enco_Air [OPPO-ENCO-AIR-WHITE] 279 EarBuds White In stock 5.2
OPPO Enco_Air3 [OPPO-ENCO-BUDS3-PURPLE] 199 EarBuds Purpule In stock 5.3
OPPO Enco_Air3 [OPPO-ENCO-BUDS3-WHITE] 199 EarBuds White In stock 5.3
OPPO Enco_Buds2 [OPPO-ENCO-BUDS2-MONLIGHT] 99 EarBuds Blue In stock 5.2
OPPO Enco_Air2 [OPPO-ENCO-AIR2-WHITE] 169 EarBuds White Incoming 5.2
OPPO Enco_Buds [OPPO-ENCO-BUDS-WHITE] 149 EarBuds White Incoming 5.2
Count of Availablity for each product
2 (33,33%)
4 (66,67%)
Availablity
In stock
Incoming
Price averege (DT)
182,33
0,00
364,67