Proyecto Web Scraping
Grupo PI ALFA_CBBCCC4

MERCADO LIBRE - Precios de Aire Acondicionado

Integrantes:

- Bustos Jonathan Nicolas
- Cappelletti Daniela
- Catalano Pérez Diana
- Cellone Franco

Links relacionados:

- **Github:** https://github.com/ispc-programador2022/CBBCCC4
- **Trello:** https://trello.com/b/GFwutGmp/proyecto-tecnol%C3%B3gico-integrador
- **M.Libre:** https://listado.mercadolibre.com.ar/aire-acondicionado

Introducción

En la actualidad, el web scraping es una técnica muy utilizada y consiste en la extracción de datos significativos de una o varias páginas web, para el análisis o manipulación de los datos, lo cual es de gran importancia para monitorear precios de la competencia, comparación de precios en tiendas, recolectar o detectar cambios en una web, analizar enlaces de una web para buscar links, etc.

Como señalamos anteriormente, el web scraping se enfoca en la transformación de los datos sin estructura (como el formato HTML) en datos estructurados que se pueden almacenar y analizar en una base de datos creada en una hoja de cálculo o cualquier fuente de almacenamiento.

Algunas limitaciones del web scraping son que al automatizar un programa para extraer datos de la web, y si esta cambia de un día para otro, es posible que dicho programa genere errores, por lo cuál se debe tener esto pendiente.

En el presente informe se mostrará la metodología seguida por el Grupo PI ALFA_CBBCCC4 en cuánto al scraping realizado a la página www.mercadolibre.com.ar en el rubro de aires acondicionados, posteriormente graficados e interpretados.

Metodología

Inicialmente, el grupo se reunió vía zoom para seleccionar el tema a escoger según la consigna dada por los profesores de la materia Proyecto Integrador. Cada integrante hizo una propuesta del tema a desarrollar y quedó seleccionado para el análisis de datos y las comparaciones de los precios, diferentes modelos y marcas de aires acondicionados publicados en Mercado Libre Argentina, teniendo en cuenta las diferentes alternativas para la efectivización de la compra.

Para iniciar el trabajo, creamos un tablero en Trello con sus respectivas tarjetas de acuerdo a la prioridad del proyecto. Y, para que todos los integrantes del grupo tuviésemos acceso a los archivos e ir subiendo información, utilizamos el repositorio de Git y Github.

Paralelamente empezamos a indagar las páginas de Mercado Libre Argentina en el rubro de aires acondicionados para conocer su estructura y las principales características, de manera que pudiésemos realizar un trabajo según la consigna dada. La plataforma de Mercado Libre (ML) muestra los productos de profesionales especializados, de pequeñas tiendas y grandes marcas a través de categorías, la publicación muestra las imágenes, precio y la descripción del producto. También los medios de pago que sean convenientes al comprador y el tipo o medio de transporte a utilizar.

Una vez entendida la estructura de la página, como se muestra en la figura 1, copiamos la URL al Jupyter y mediante un import request hicimos la solicitud de tipo Get para que la página devolviera el código HTML. Seguidamente, importamos la librería bs4 BeautifulSoup

para posteriormente hacer el scraping del HTML e ir elemento por elemento de lo que deseábamos encontrar en la página de ML. Se hizo la respectiva comprobación con status_code para corroborar que el procedimiento, hasta el momento, estuviera correcto.

Figura 1. URL e import request del HTML de ML.

Luego se definen varias variables para que al buscar por consola / teclado nos arroje el contenido que buscamos. Estás variables mediante el uso de xpath nos permitieron contener todos los productos que deseábamos (nombre y precio del producto con su respectiva URL). Además, importamos la librería Pandas que nos permite exportar los datos requeridos y guardarlos en formato CSV. Se realizó el scraping a 40 páginas con un contenido de 50 productos cada una.

El procesamiento de los datos se realizó mediante R, lo que nos permitió obtener las tablas con información referida a “tipos de aires acondicionados según título del artículo” y el “precio promedio de los aires acondicionados según tipo” Para finalizar usamos hojas de cálculo de google para diseñar los gráficos queso pueden ver en la Figura 2.

Para el análisis de los datos consideramos las variables tipo de aire acondicionado según la publicación y precio promedio según el título del artículo. Para lo cual utilizamos herramientas estadísticas que permitieron realizar una comparativa de precios y de los tipos de aires acondicionados sin considerar marca o modelo porque no lo consideramos relevante.

Figura 2. Código utilizado para el procesamiento de los datos.

```r
#Cargamos las librerías

library(tidyverse)

#Descargo la bbss

meli <- read_csv("listado_mercado_libre_.csv")

glimpse(meli)

meli %>% count("Articulos")

#Voy a limpiar toda la columna de artículo

library(janitor)
meli$Articulos <- make_clean_names(meli$Articulos)

library("stringr")

# consider a string "Hello Geek"
# replace the character 'e' in  "Hello Geek"
# with "E"

meli$Articulos <- str_replace_all( meli$Articulos,"_", " ")

#Voy a quedarme solamente con aquellos articulos que dicen calor

aire_caliente <- meli %>%
  filter(str_detect(Articulos, "calor")) %>%
  filter(!str_detect(Articulos, "frio"))

aire_frio <- meli %>%
  filter(str_detect(Articulos, "frio")) %>%
  filter(!str_detect(Articulos, "calor"))


aire_frio_calor <- meli %>%
  filter(str_detect(Articulos, "frio")) %>%
  filter(str_detect(Articulos, "calor"))

aire_ni_ni <- meli %>%
  filter(!str_detect(Articulos, "frio")) %>%
  filter(!str_detect(Articulos, "calor"))

aire_split <- meli %>%
  filter(str_detect(Articulos, "split"))

aire_inverter <- meli %>%
  filter(str_detect(Articulos, "inverter"))

aire_no_inverter <- meli %>%
  filter(!str_detect(Articulos, "inverter"))

aires <- rbind(aire_caliente %>%
                 mutate(Tipo = "Caliente"),
               aire_frio %>%
                 mutate(Tipo = "Frio"),
               aire_frio_calor %>%
                 mutate(Tipo = "Frio y Calor"),
               aire_ni_ni %>%
                 mutate(Tipo = "Sin aclaración"))

aires_inverter <- rbind(aire_inverter %>%
                          mutate(Tipo = "Inverter"),
                        aire_no_inverter %>%
                          mutate(Tipo = "No Inverter"))

#Quiero saber el precio promedio según el tipo

aires$Precios = as.integer(aires$Precios)

precio_tipo <- aires_inverter %>%
  mutate(Precios = as.integer(Precios)) %>%
  na.omit(Precios) %>%
  group_by(Tipo) %>%
  summarise_at(vars(Precios), list(name = mean))

max(aire_inverter$Precios)

aire_inverter %>%
  na.omit(Precios) %>%
  mean(Precios, na.rm = TRUE)

#Frecuencia de palabras en los títulos de los artículos

my_stopwords <- read_csv("~/Documents/Ciencia de datos/R-Projects/Clases/my_stopwords.csv")

word_freq <- aires %>% unnest_tokens(word, Articulos, drop = TRUE) %>%
  anti_join(my_stopwords) %>%
  count(word, sort = TRUE)%>%
  top_n(20)
```

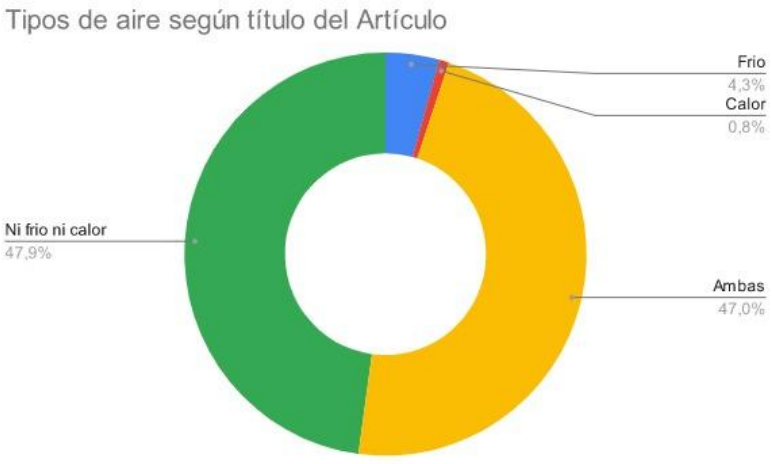
Análisis de los datos obtenidos de ML

1) Tipos de aires acondicionados según título del artículo

Estos datos fueron de los más sencillos, como se señaló el scraping se realizó en 40 páginas con un total de 50 productos por página. Como puede observarse en la tabla 1 se obtuvieron 4 tipos de aires acondicionados según el título del artículo ofrecido, frío (4.3%), calor (0.8%), frío-calor (47.9%) y ambos (47%). Posiblemente estos resultados obedezcan al cambio de estación en Argentina en el que aún el calor no es muy fuerte (Considerar que al momento del scraping aún no estamos dentro del período estival).

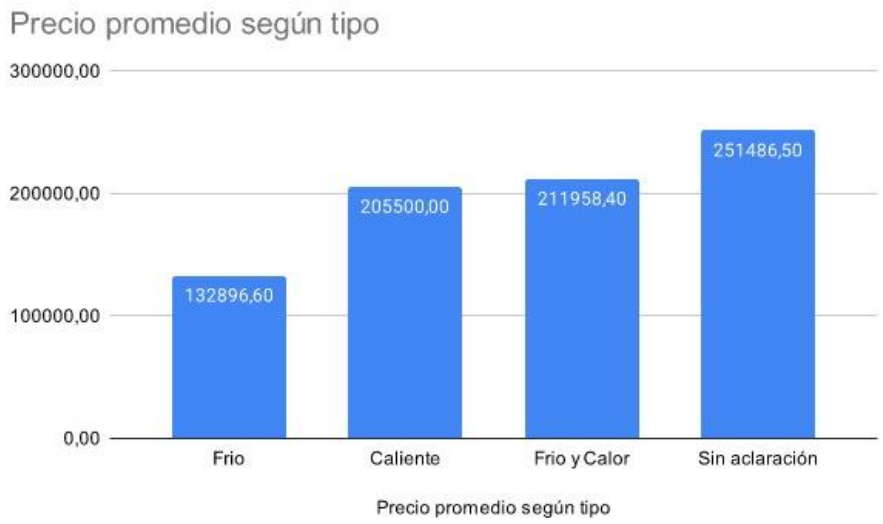
Tabla 1. Tipos de aire acondicionado según título del artículo

Tipo aire acondicionado	Según título del artículo (%)
Frío	4.3
Calor	0.8
Ni frío ni calor	47.9
Ambos	47.0
Total	**100.0**



2) Precio promedio de los aires acondicionados según tipo

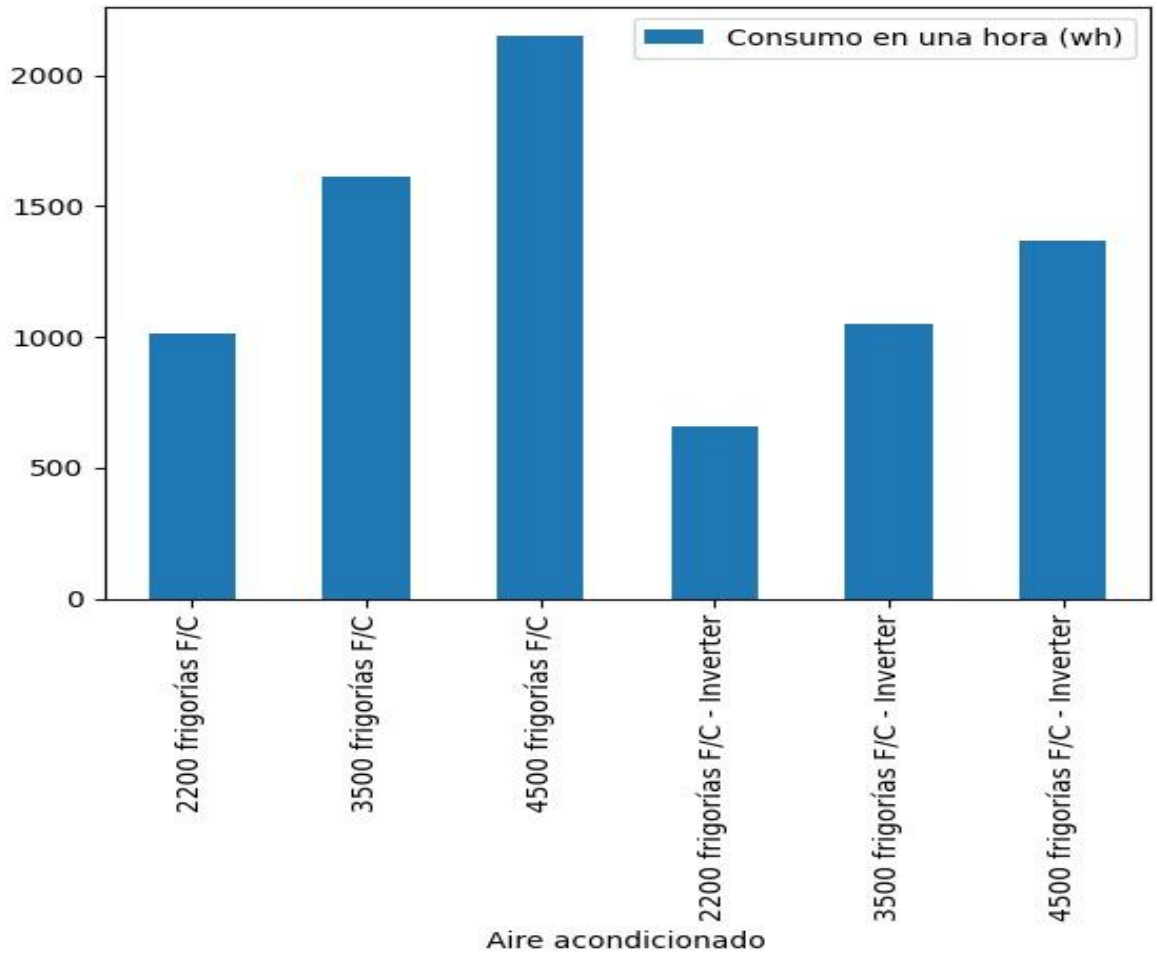
Según el scraping realizado a las páginas de productos de aires acondicionados de ML, el precio de estos equipos estaba ubicado en un rango entre $132.000 y $300.000, siendo la categoría Frío, la más baja en precios promedio de $ 132.896 y los aires de un mayor precio promedio no contaban con dicha aclaración en el título de la publicación.

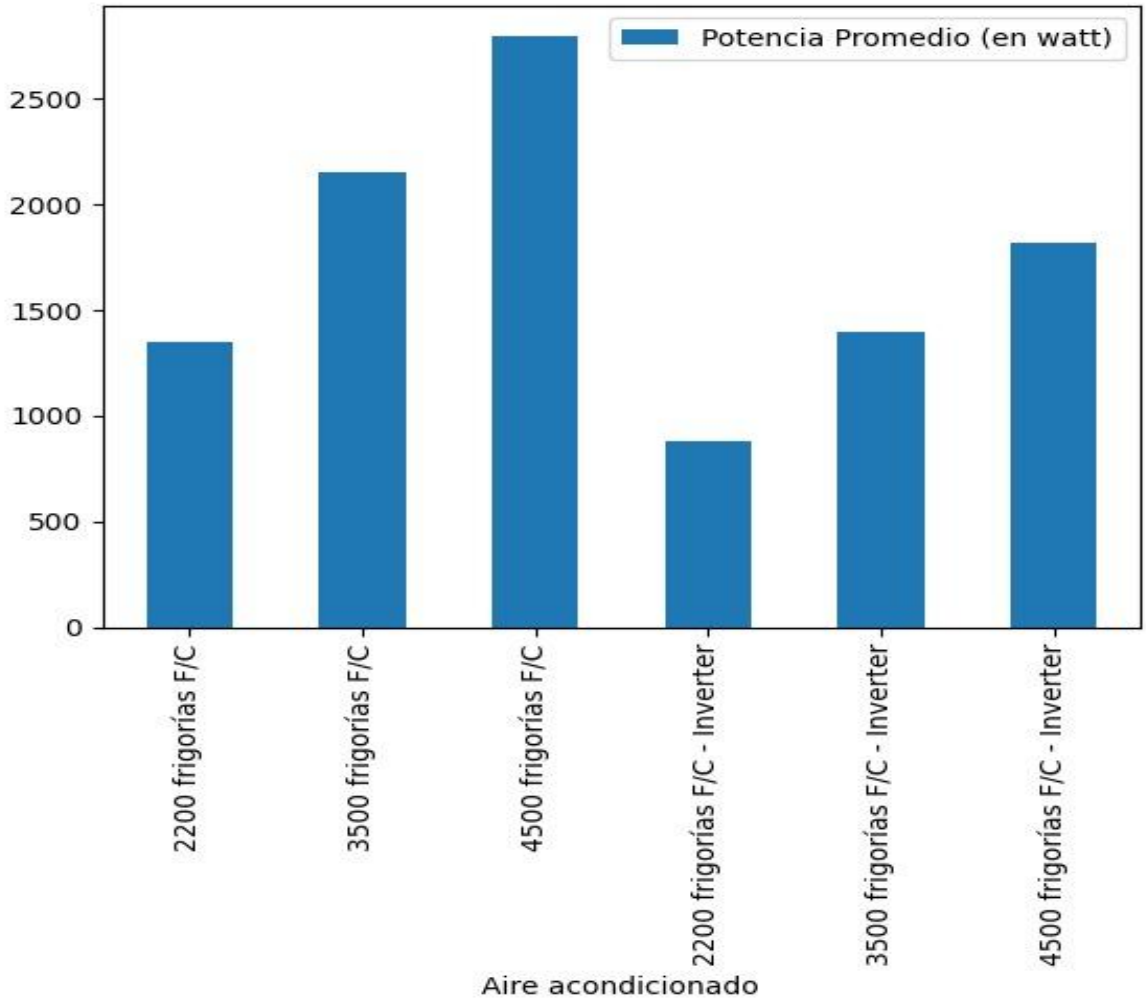


CONCLUSIONES

A la hora de adquirir un equipo frío / calor hay que tener en cuenta no solamente el precio sino también el consumo promedio que tendrá el mismo, ya que este impactará en la factura de electricidad, es por ello que es imperativo leer la etiqueta de eficiencia energética, definir qué tipo de tecnología es más conveniente y el tamaño del espacio donde se colocará el aparato.

Razón por la que, nos pareció importante considerar también, una pequeña información obtenida de la página: https://www.argentina.gob.ar/enre/uso-eficiente-y-seguro/consumo-basico-electrodomesticos, en donde se obtienen los consumos básicos promedio, según la potencia y tipo de tecnología, que no debería de ser un dato menor a la hora de elegir cuál es la mejor opción de compra en función de los costos y beneficios.




Potencia Promedio (en watt)
2500
2000
1500
1000
500
0
2200 frigorías F/C
3500 frigorías F/C
4500 frigorías F/C
2200 frigorías F/C - Inverter
3500 frigorías F/C - Inverter
4500 frigorías F/C - Inverter
Aire acondicionado