Trabajo práctico Ciencia de Datos

Web scrapping

Introducción

¿Qué es el web scrapping?

Definición: técnica que permite recolectar grandes volúmenes de datos directamente desde páginas web de forma programática y estructurada.

Ventajas:

- **Eficiencia:** Permite recolectar grandes volúmenes de datos de forma rápida y eficiente, en una fracción del tiempo que tomaría hacerlo manualmente.
- **Estructuración de Datos:** Transforma información desorganizada de la web en formatos limpios y analizables (como tablas o bases de datos).
- **Monitoreo y Análisis:** Útil para monitorear precios, analizar mercados, recopilar noticias, investigar la competencia, y seguir cambios constantes en la web.
- **Toma de Decisiones:** Centraliza y organiza la información para mejorar la capacidad de análisis y, por ende, la calidad de las decisiones.

Aplicación del web scrapping al TP

Expectativa

Objetivos:

- Recolectar información de las páginas Yenny y Buscalibre
- Hacer un resumen de los precios de cada página
- Comparar los precios de cada página
- Hacer un gráfico donde se vea claramente esta comparación

Dificultades esperadas:

- Buscar una forma de encontrar los mismos libros en las distintas páginas, a fin de poder comparar
- Manipular la información de forma ordenada
- Hacer los gráficos de manera profesional

Realidad

Resultados:

- La comparación de libros es muy complicada, categorías distintas a través de distintas paginas
- Tampoco tienen la opción de sortear por autor adecuadamente
- Disparidad en la cantidad de libros que tienen las paginas haciendo el análisis menos adecuado

Solución planteada: pasar a celulares

- Menos diversidad
- Prácticamente nula diferenciación por categorías
- Mayor facilidad al comparar por marcas y modelos específicos

Nuevo objetivo:

- Comparar las tiendas de Claro y Personal
- Ver qué pagina tiene mejores precios
- Presentar la información de manera tal que resulte fácilmente entendible

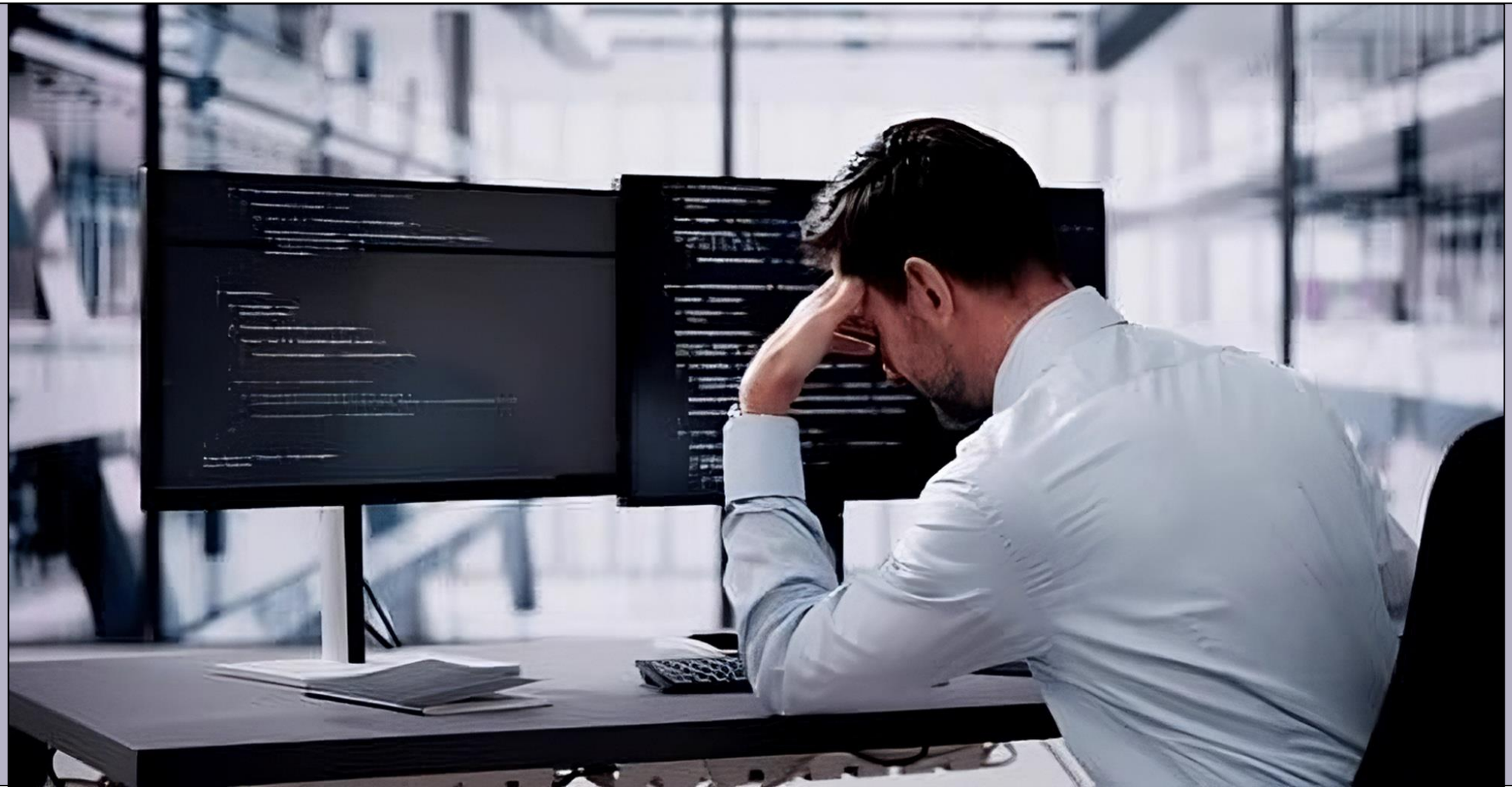
Dificultades esperadas:

- Inconsistencias a la hora de nombrar los productos
- Irregularidades en los precios o modelos duplicados
- Estructura del código de las paginas web (por ej, uso poco entendible de las clases)

¿Qué hice entonces?

- Hice un código para levantar datos de Personal y Claro por separado .
- Al no tener Claro páginas, levanté todos los celulares de la página y una cantidad equivalente de la tienda de Personal (4 de las 8 páginas).
- Hice primero el código de Personal y después lo modifique para Claro, por eso es ultimo usa Selenium a pesar de ser innecesario (me ahorra tiempo).

Dificultades



A la hora de scrapping

Personal

- Identificar elementos correctos.
- Sortear inconsistencias en el armado de las clases de la pagina.
- Conformar el for loop de manera eficiente.
- Determinar como levantar la información.

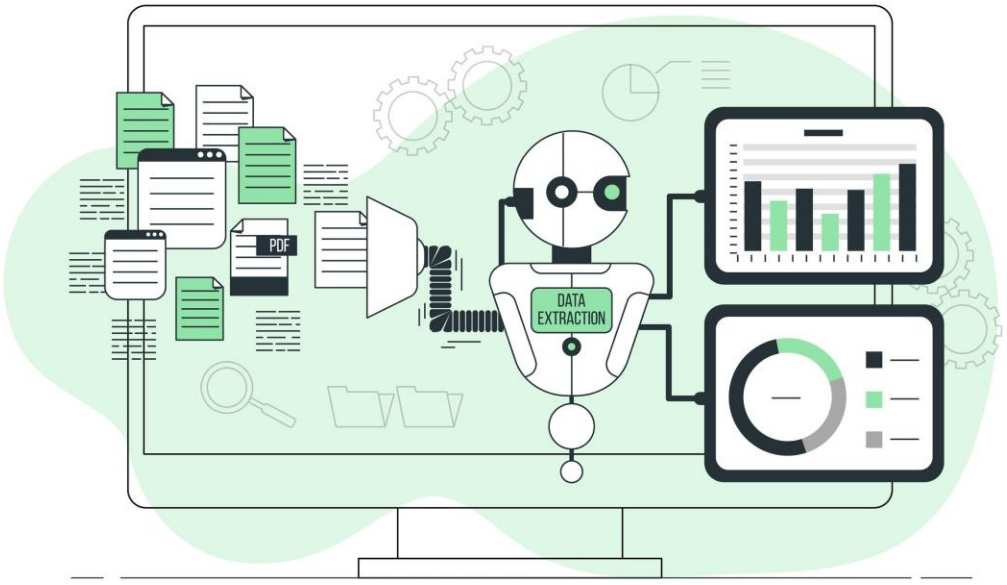
Claro

- Extracción de links, usaban la misma clase para todos los links sean de celulares o de categorías, etc.
- Problemas a la hora de usar el xpath y el css de manera que recolecte la información que yo quería

A la hora de graficar la información

- Dificultad en comprender cuales son los datos que hay que comparar
- Elegir cómo presentar la información seleccionada entre todas las opciones que hay
- Darle la claridad visual necesaria a los gráficos
- Como trabajar con ggplot

Extracción de datos



Los datos extraídos fueron

MARCA DE CELULAR (SAMSUNG, MOTOROLA, APPLE, ETC)

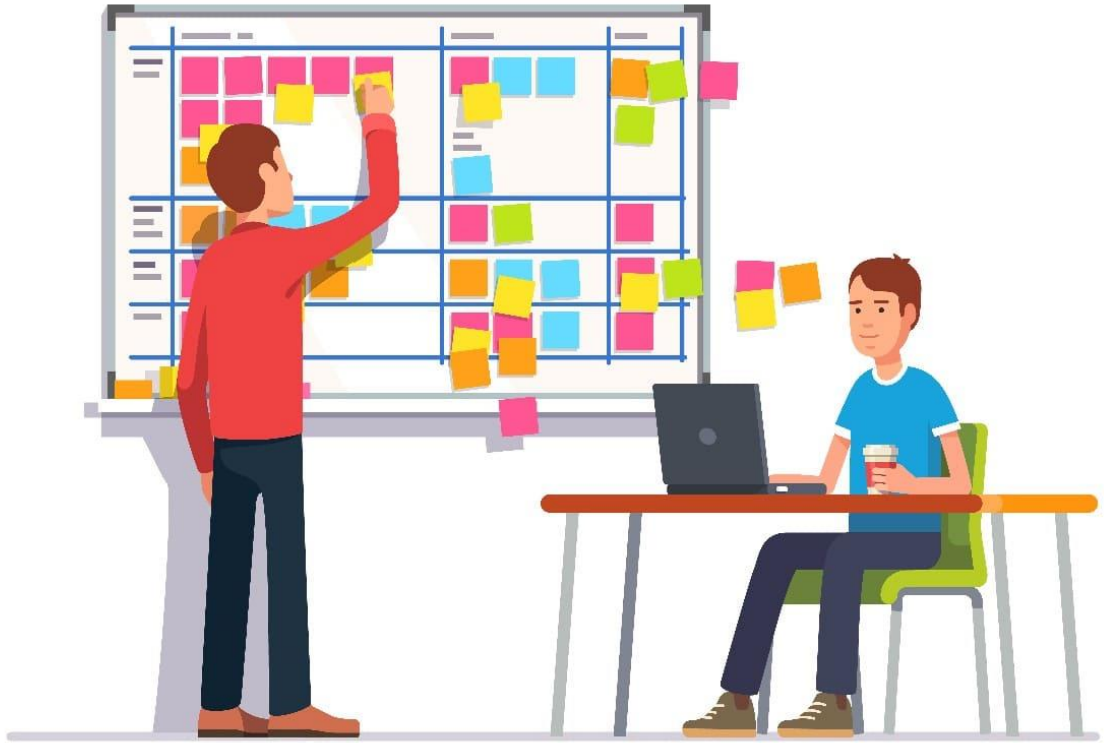
MODELO

PRECIO DE LOS DISPOSITIVOS

MÉTODO DE PAGO (CUOTAS CON TARJETA, CONTADO, ETC)

URL

Metodología aplicada



Metodología de herramientas

Selenium para analizar todas las páginas de Personal y extraer el HTML de cada una

Rvest para levantar los datos de los HTML extraídos

Tidyverse para la manipulación y estructuración de los datos

Metodología de trabajo

Extraje los datos: nombre, precio, método de pago, link y “todo” (que quedó de la clase de scrapping)

Dividí el nombre para sacar la marca y el modelo en columnas separadas

Limpié el precio para dejarlo como “numeric” en R

Simplifiqué y blanquee los métodos de pago

Normalicé los nombres y las marcas (por ejemplo en uno en vez de Apple decía iPhone) para después poder unirlos

Análisis Exploratorio de Datos

Visión general de los datos

Volumen de datos:

- De Personal pude recolectar 83 datos de celular, los cuales 3 tuve que cortar porque se repetían (no me pregunten por qué)
- De claro pude levantar los 43 celulares que había en la página

Columnas/variables:

- Columnas al principio: nombre, precio, met_pago, link, todo
- Columnas al final: marca, modelo, precio_claro/personal, dif_precio, mas_económico, met_pago_claro/personal, etc

Limpieza y tratamiento

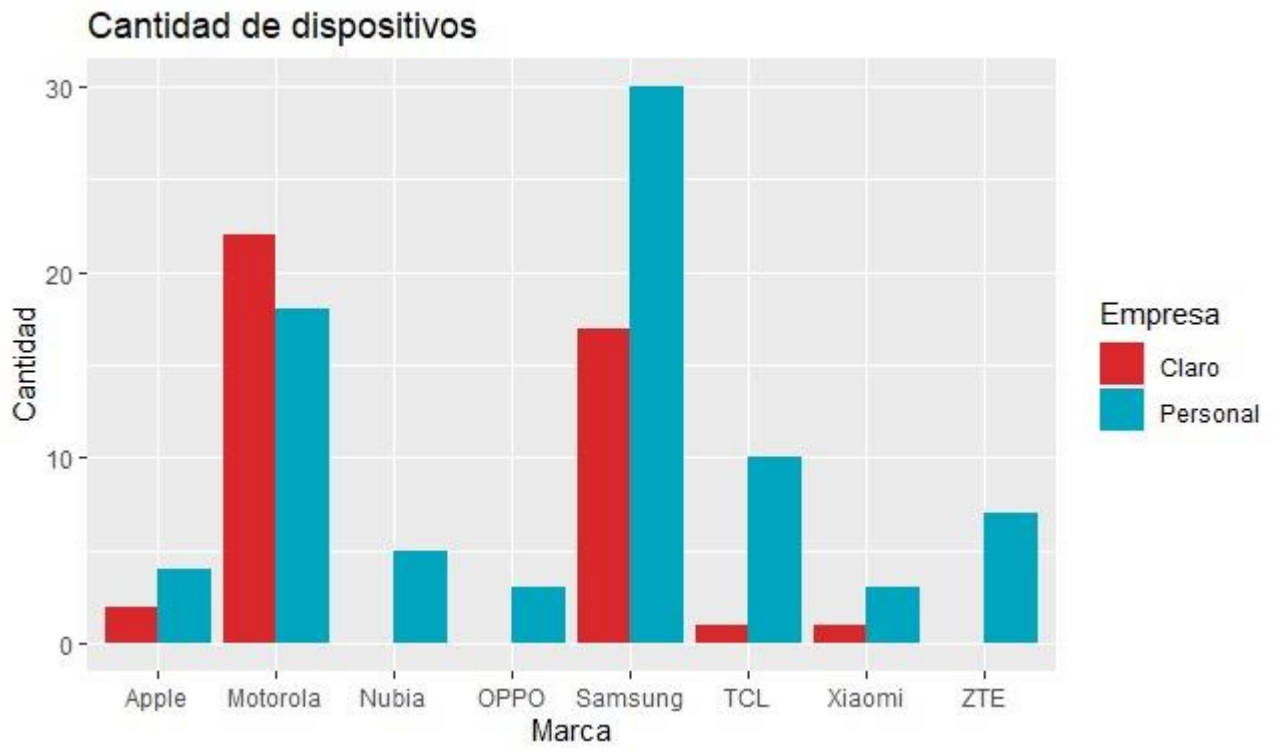
- **Valores faltantes**: Los celulares que estaban en una página y en otra no los mantuve para el volumen total pero los saqué al hacer la comparación directa
- **Tipos de dato**: Pasar los precios de character a numeric
- **Duplicados**: En personal habían 3 casos de celulares que se repetían pero con distinto precio, me quedé con el más barato
- **Estandarización de Nombres:** Saqué todo lo que sobraba en los modelos y marcas para que no genere problemas al comparar

Análisis descriptivo

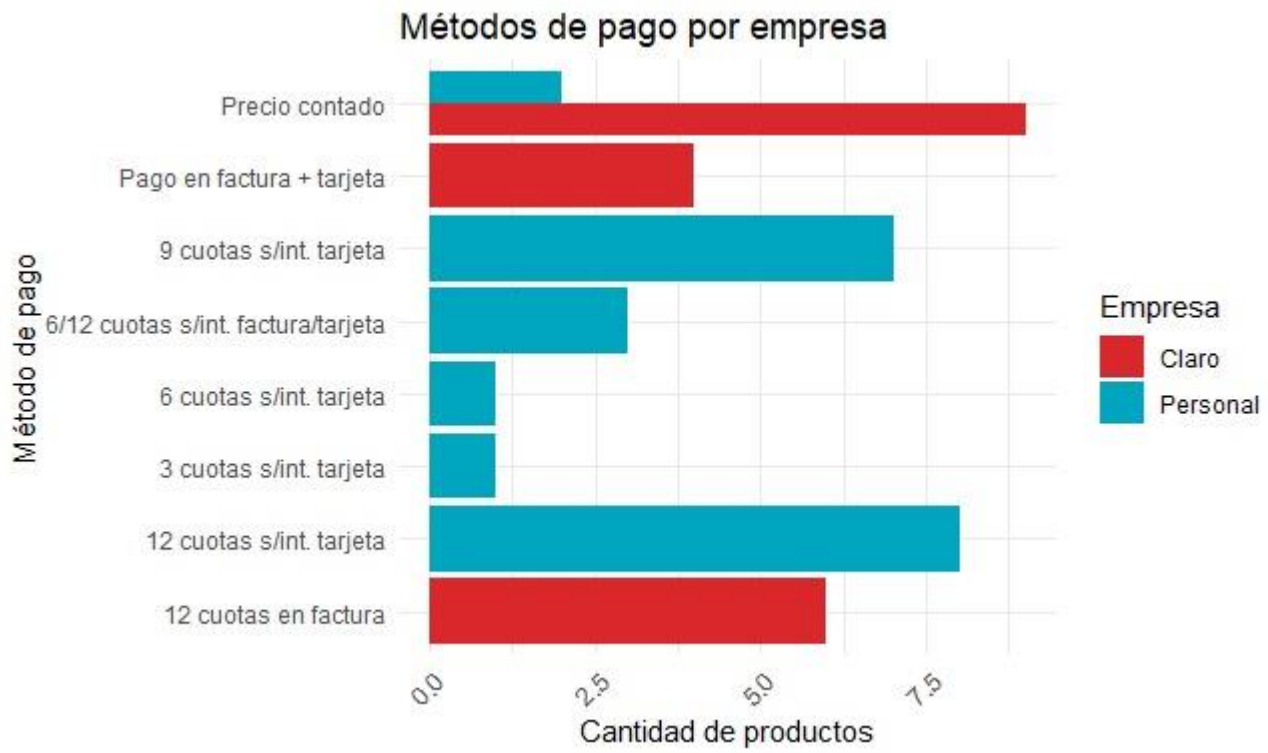
- **Distribución de precios**: Usé un boxplot para mostrar la distribución de los precios de cada tienda, esto se complementa con el resto de datos
- **Variedad**: Quería mostrar con un gráfico de barras como Personal tiene bastante más variedad que Claro
- **Método de pago por empresa**: A demás de tener más variedad, personal tiene más opciones de cuotas y pagos

Resultados

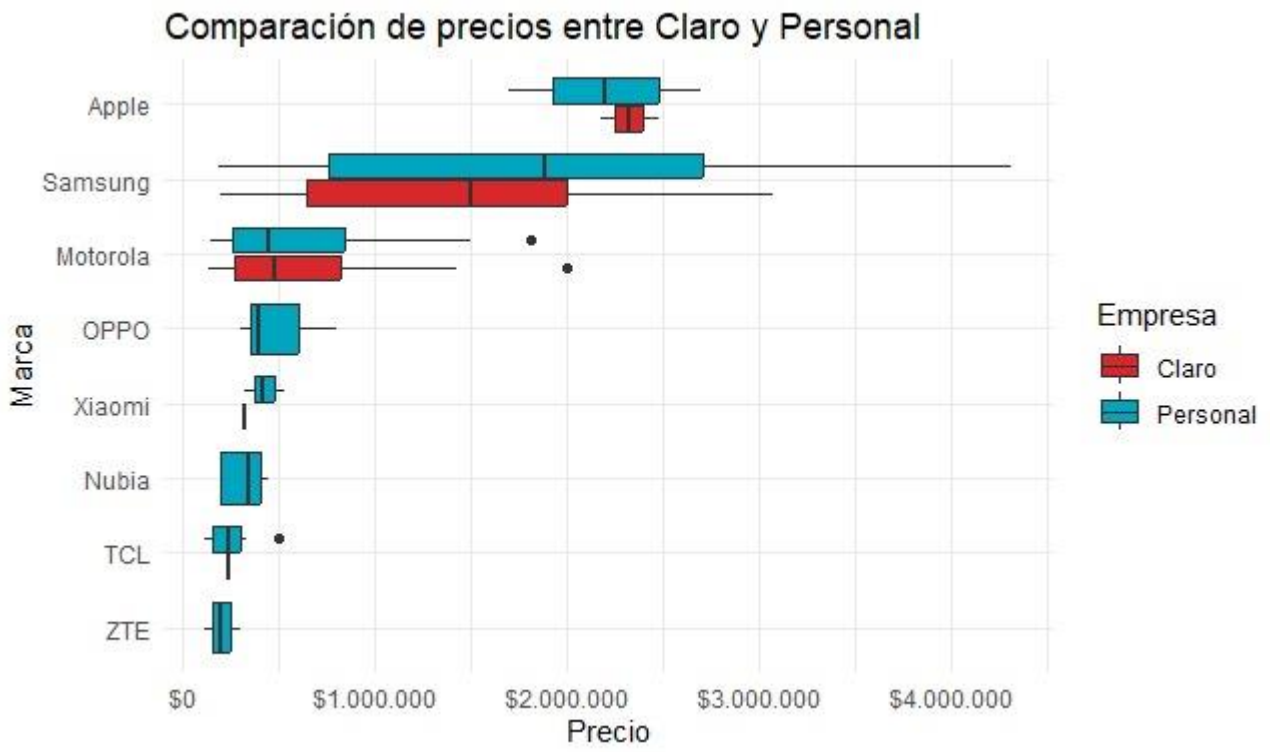
Personal tiene mayor variedad que Claro



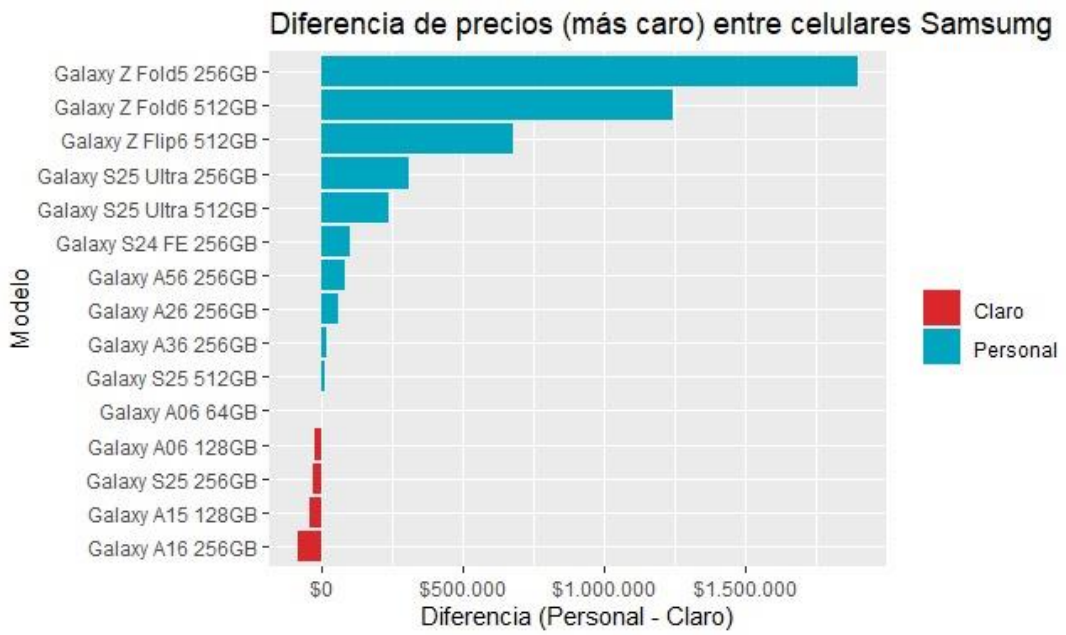
Personal tiene más opciones de pago que Claro

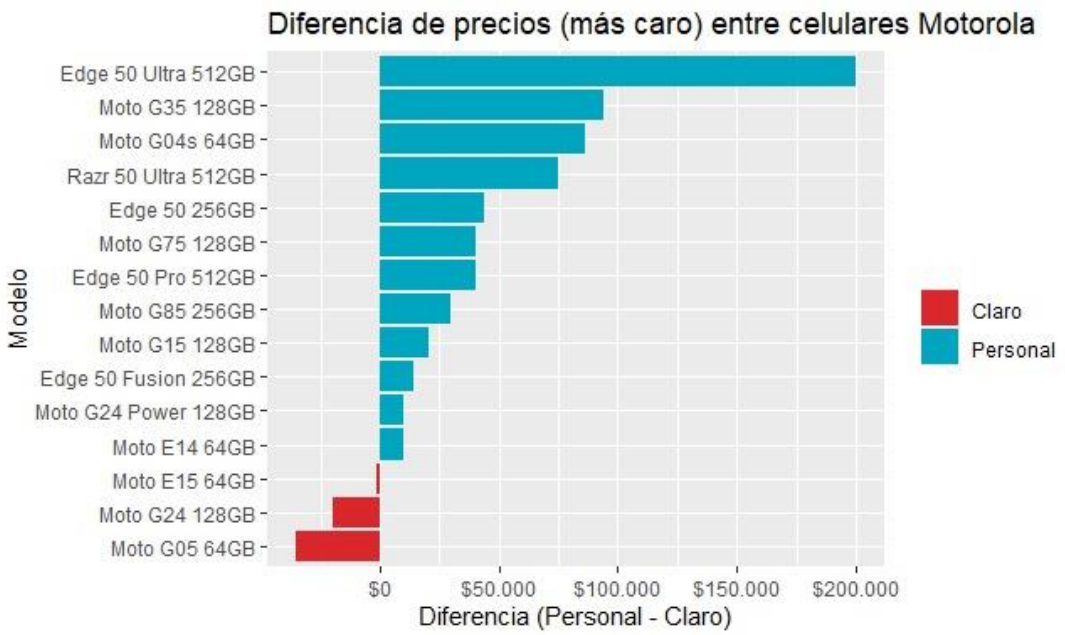


Sin embargo, Claro presenta precios más bajos en los modelos comparables.

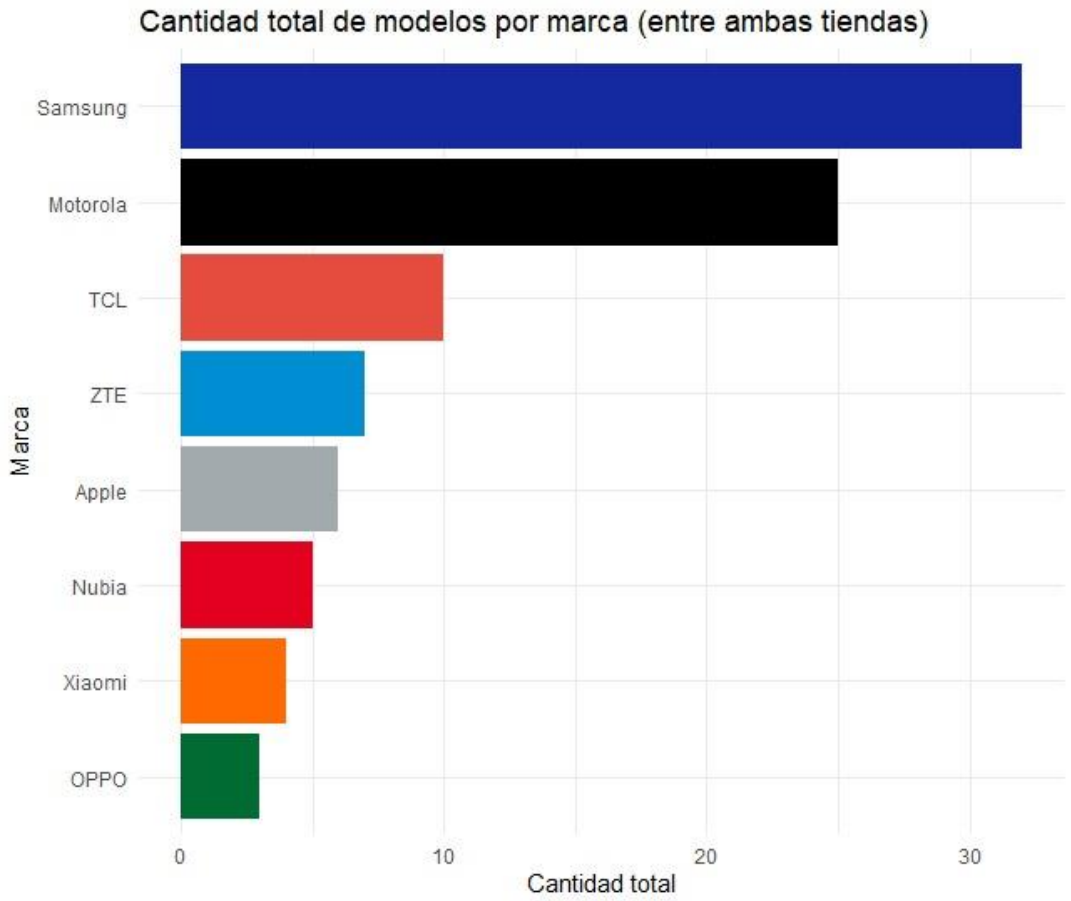


Las diferencias de precios son mas notorias en gamas altas y medias





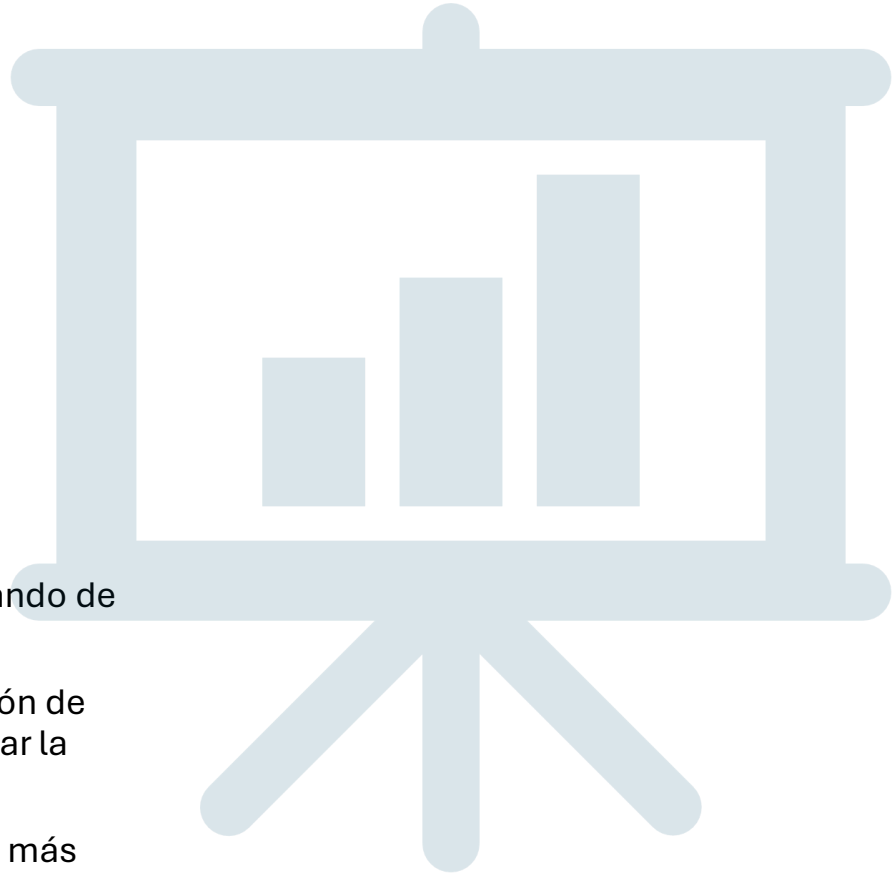
Samsung y Motorola son las opciones más populares



El por qué de los gráficos seleccionados

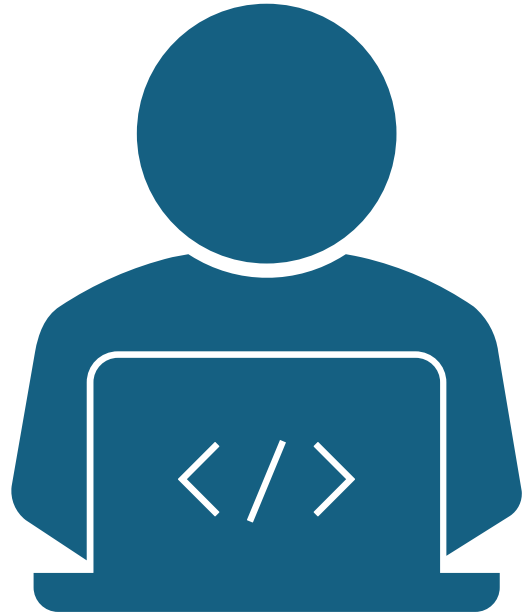
¿Por qué los gráficos seleccionados?

- Usé principalmente gráficos de barras porque, al estar hablando de cantidades, es más intuitivo que por ejemplo, un pie chart
- En el gráfico del boxplot, quería mostrar no solo la distribución de precios entre tienda y tienda si no que también quería mostrar la distribución entre marcas
- También hice dos gráficos de comparación entre las marcas más representativas porque no tenía suficiente información para las demás



Ventajas del tablero dinámico Shiny

- Hace el transmitir información un proceso más interactivo
- Obliga a el analista a profundizar más en su comprensión de los datos que está exponiendo para poder transmitirlos dinámicamente sin problemas de entendimiento
- Se ve mucho más profesional
- Ayuda a refinar las habilidades a la hora de programar



Limitaciones y trabajo futuro

Limitaciones

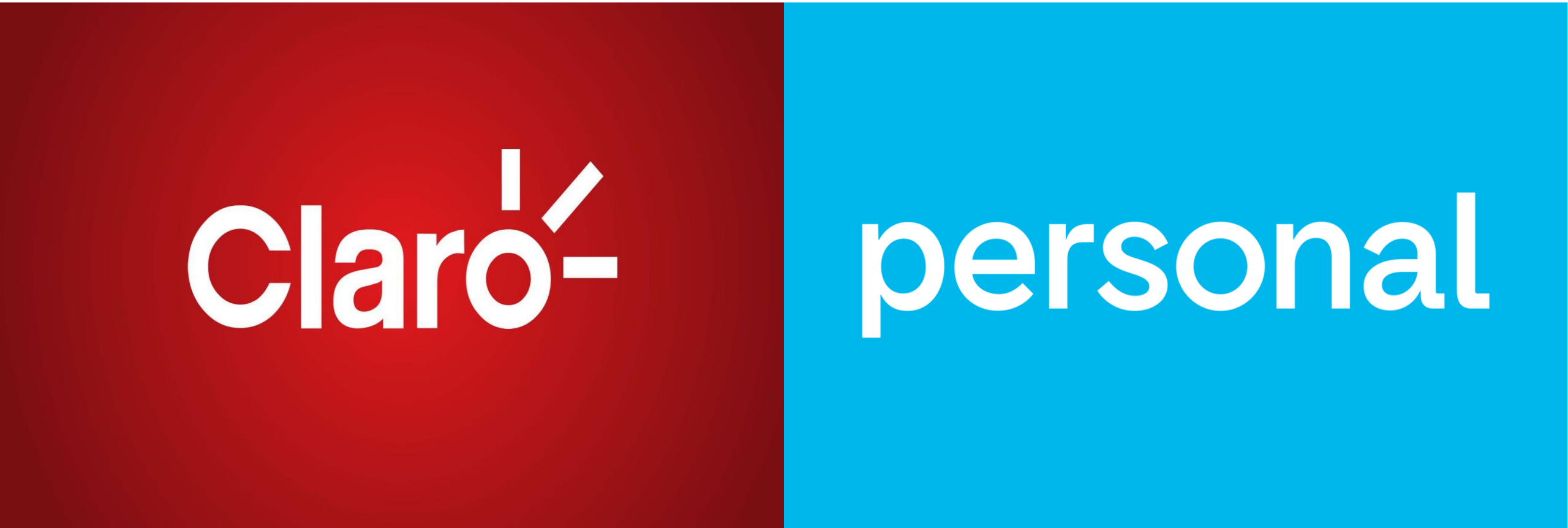
- No pude desarrollar la aplicación del Shiny como me hubiese gustado, me gustaría darle más personalización
- No pude levantar toda la cantidad de datos que quería, me hubiese gustado agregar aún más páginas
- El HTML de las páginas cambia bastante, tuve que agregar un botón anti popups porque metieron uno que no aparecía antes

Trabajo futuro

- Agregar un mayor número de páginas para tener una mayor representación de popularidad de marcas y distribución de precios
- Trabajar en la presentación visual de la aplicación Shiny
- Buscar otro tipo de categorías las cuales comparar para aprovechar mejor los datos
- Agregar medición de stock
- Hacer en análisis esporádicamente para tener una progresión de los precios y modelos vendidos

Conclusiones

Claro es más barato mientras que Personal tiene más opciones de celulares para elegir



Además...

Las marcas con más presencia son Samsung y Motorola

SAMSUNG

MOTOROLA

¡Gracias por llegar hasta acá!


That's all Folks!