

WEB SCRAPING

Pre-presentación Proyecto Grupo 2 PRODUCTO 6

Integrantes:

- FABIAN OMAR FIGUEROA BUENO
- FELIPE GUAICO BARRAZA
- DAVID VALENZUELA PEREZ

Agenda

Tecnología a utilizar
Acciones de navegación
Diagrama del proceso / Pseudo código
Estructura archivo / Output / Docs

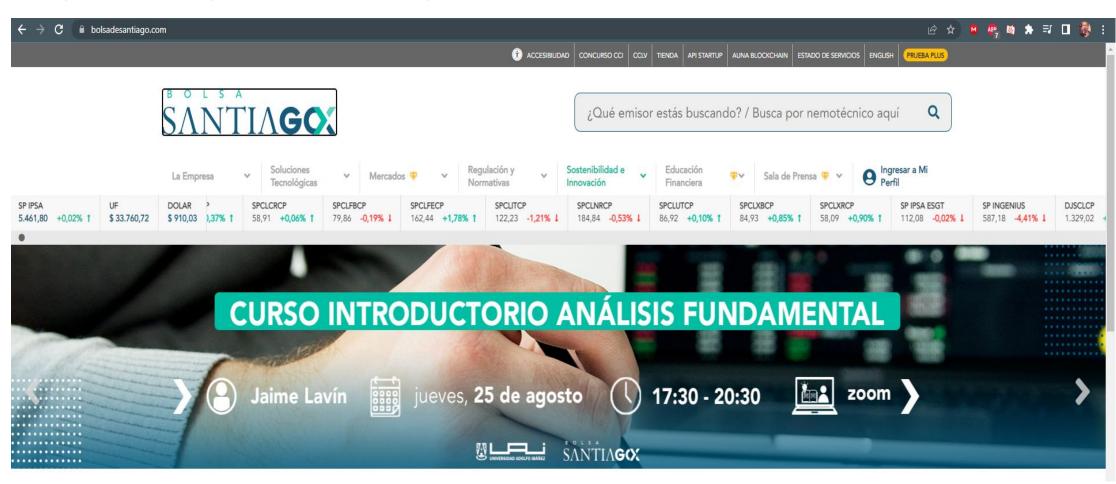
Tecnología a utilizar Proyecto – Grupo 2

Se busca que el proyecto integre todas las tecnologías vistas en clase, para lo cual, desarrollaremos un script que pueda:

- 1. Navegar la página de la bolsa de Santiago (https://bolsadesantiago.com/) con **Selenium**.
 - Todo esto con iteradores y listas para consultar un número controlado de empresas.
- 2. Descargar tablas Excel que posee la página y almacenar en una carpeta específica para cada empresa (Mediante clics con Selenium)
- 3. Descargar información adicional que se refleje en la pagina, con Beautiful Soup 4
- 4. Sobre las lógicas:
 - Idealmente si la Empresa es "nueva" en el proceso de consultas, deberemos crear una carpeta con Selenium y/o OS (Siempre y cuando exista dentro del consultador la bolsa de Santiago.
 - De lo contrario, deberá leerse la carpeta respectiva con los archivos descargados anteriormente.
 - Los puntos a y b de este apartado deberán ser controlados y manipulados con Pandas

Queremos ver la posibilidad de complementar la extracción de la información con el uso de alguna API que pueda existir para la bolsa de Santiago o páginas asociadas que puedan tener más historia para la información.

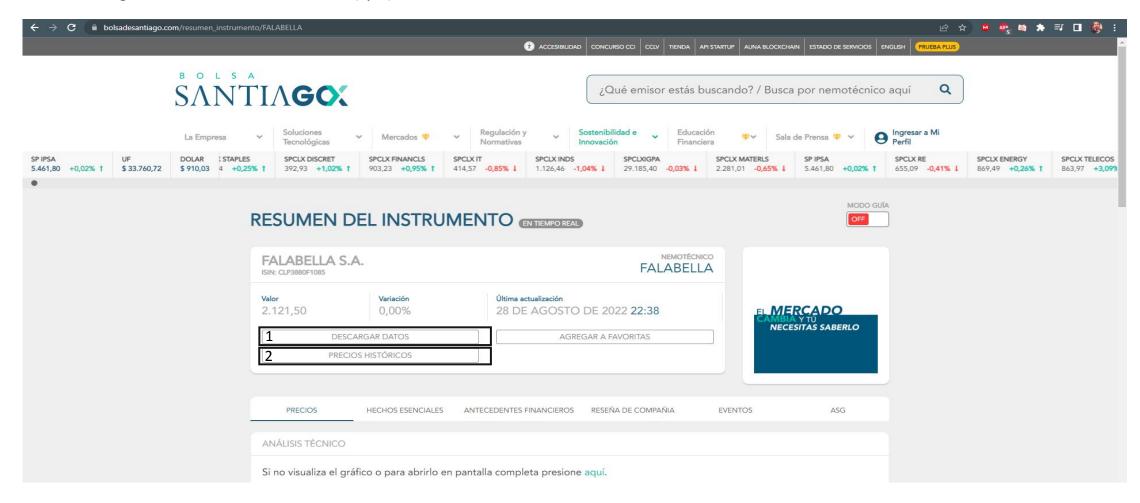
 Extracción de Información de Acciones Navegación: Ingresamos a la Página de la Bolsa de Santiago



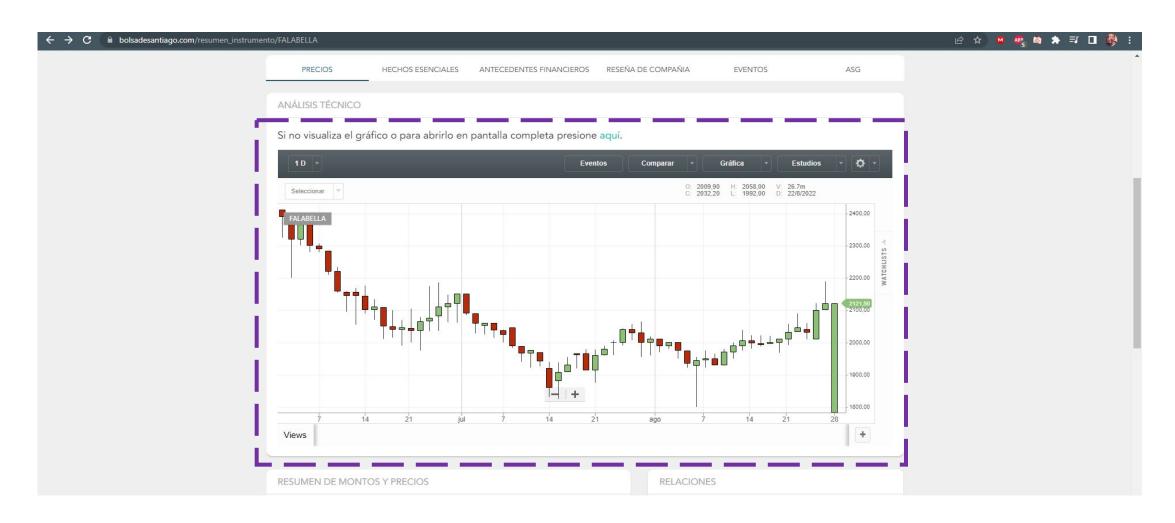
Consultamos una de las Empresas a monitorear



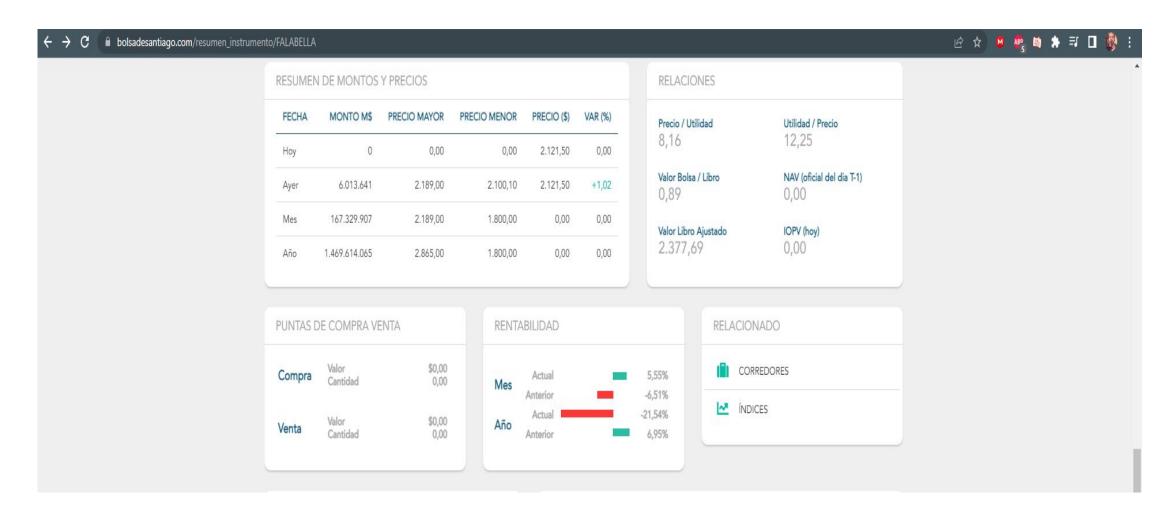
- 3. Si la empresa existe, apretar Enter e ingresar a la misma
- 4. Descargamos los Excel de los Puntos 1) y 2)



5. Abajo, aparece un gráfico que tiene más data histórica, ver si se puede conseguir con una API. – Deseable



6. En la misma página, seguimos bajando y se observan los siguientes indicadores:



7. Descargar con BS4, las últimas variaciones de Capital y respaldar en algún lado:

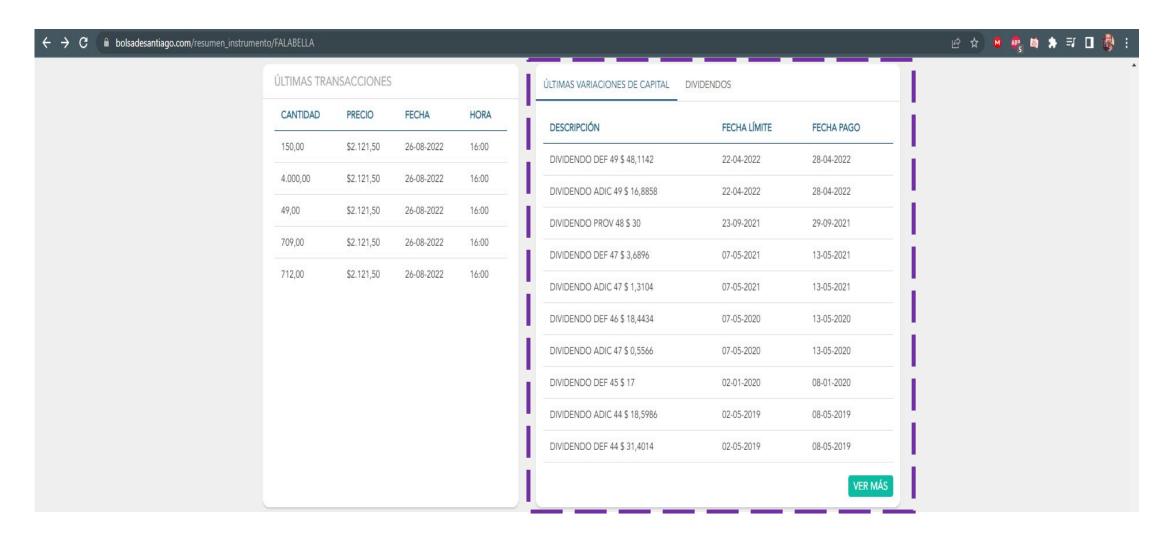


Diagrama del proceso / Pseudo código Proyecto – Grupo 2

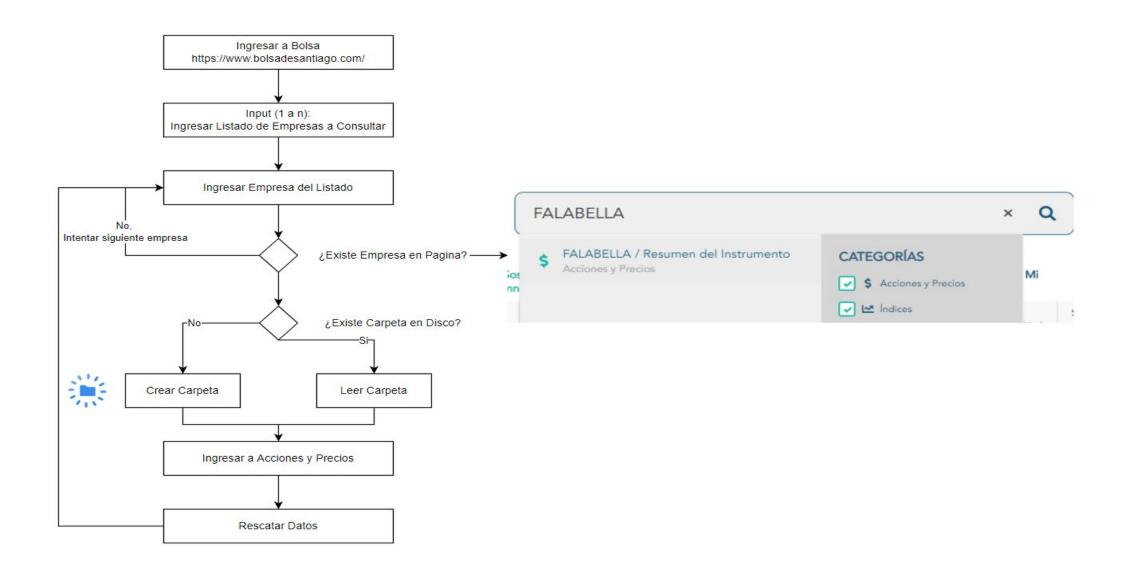
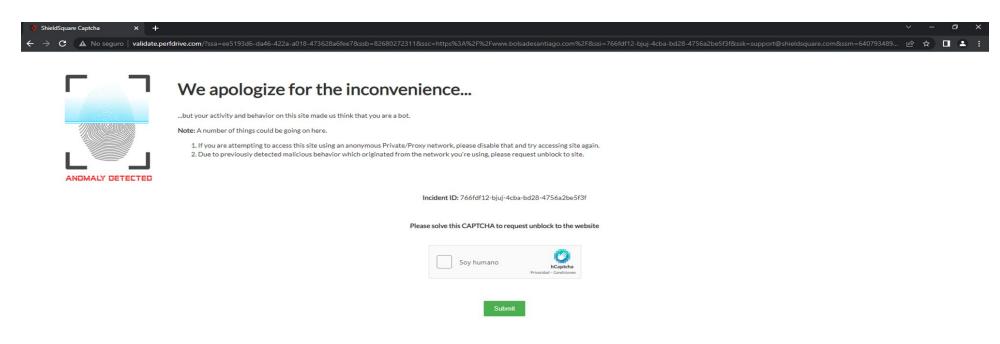


Diagrama del proceso / Pseudo código Proyecto – Grupo 2

Pero...

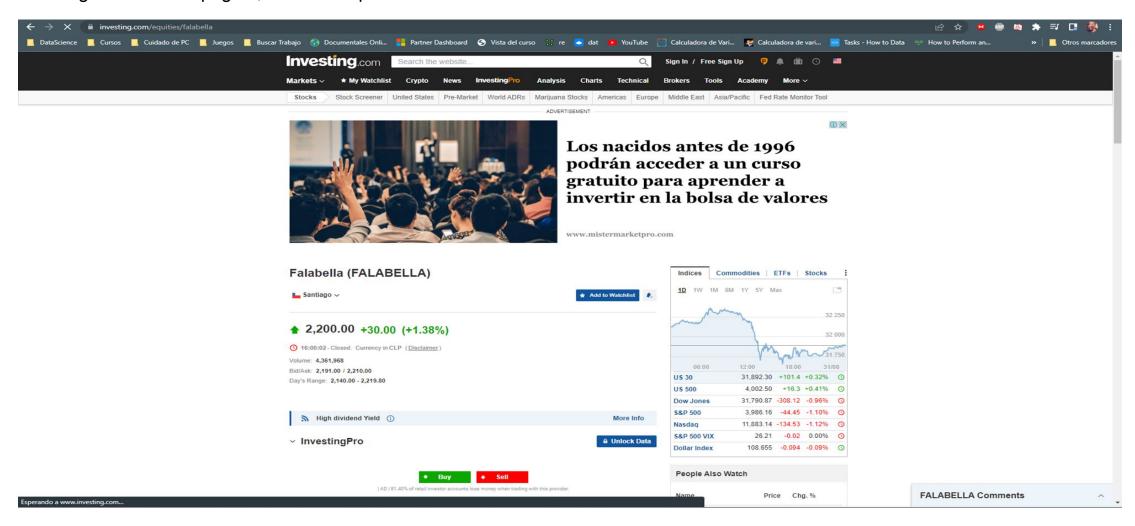


Tecnología a utilizar (Opción 2) Proyecto – Grupo 2

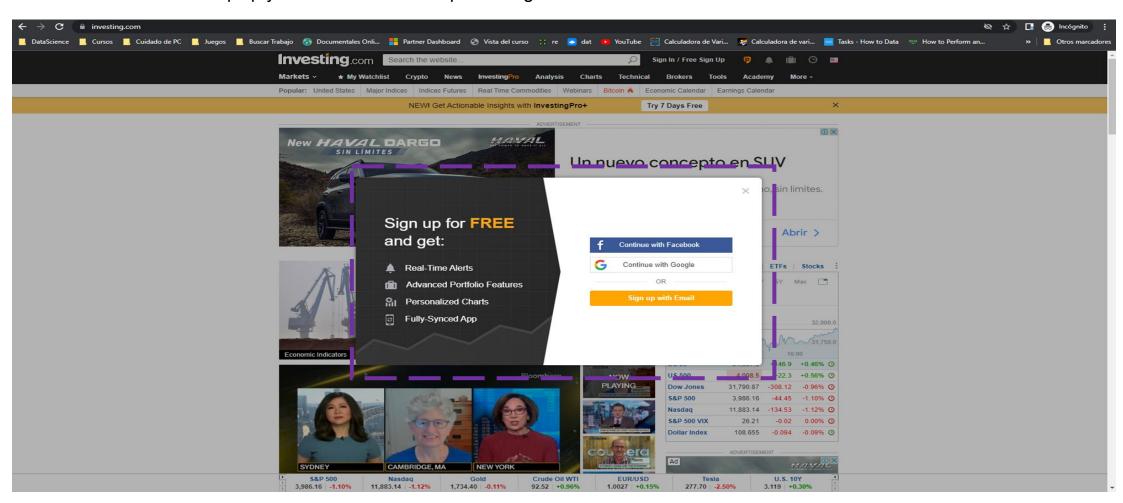
Se busca que el proyecto integre todas las tecnologías vistas en clase, para lo cual, desarrollaremos un script que pueda:

- 1. Navegar la página de la bolsa de Santiago (https://www.investing.com/equities/falabella) con Selenium.
 - Todo esto con iteradores y listas para consultar un número controlado de empresas.
- 2. Descargar información que se refleje en la página, con Beautiful Soup 4
- 3. Sobre las lógicas:
 - Idealmente si la Empresa es "nueva" en el proceso de consultas, deberemos crear una carpeta con **Selenium** y/o **OS** (Siempre y cuando exista dentro del consultador la bolsa de Santiago.
 - De lo contrario, deberá leerse la carpeta respectiva con los archivos descargados anteriormente.
 - Los puntos a y b de este apartado deberán ser controlados y manipulados con Pandas

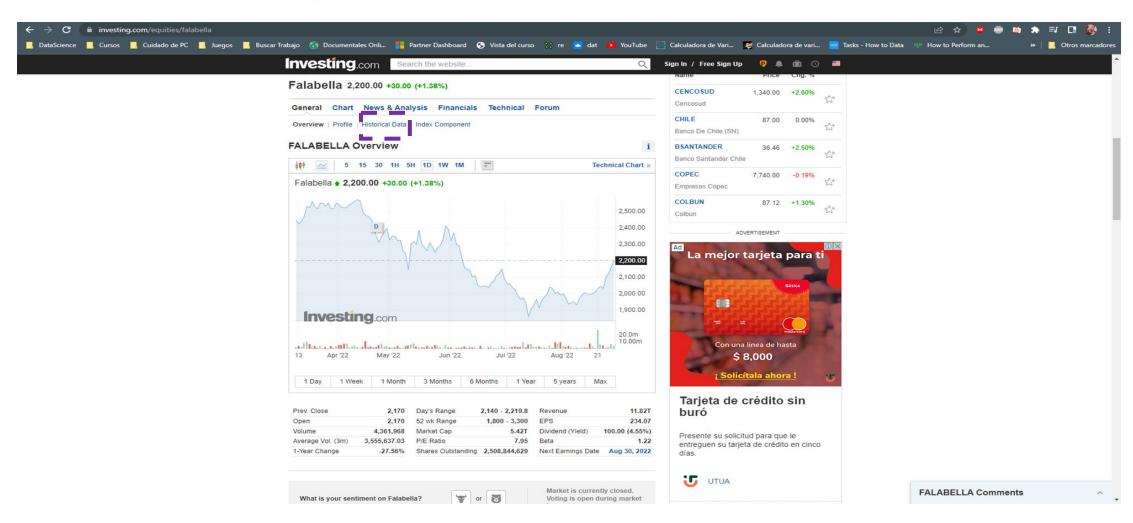
Ingresamos a la página, enfocados por ahora en Falabella.



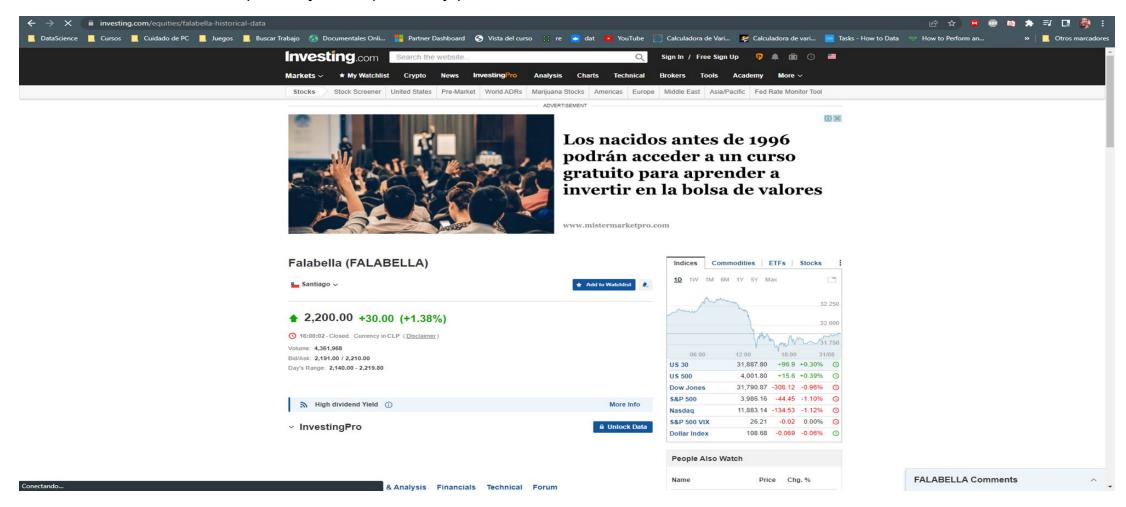
2. Controlaremos el PopUp y lo cerraremos si es que se llega a abrir:



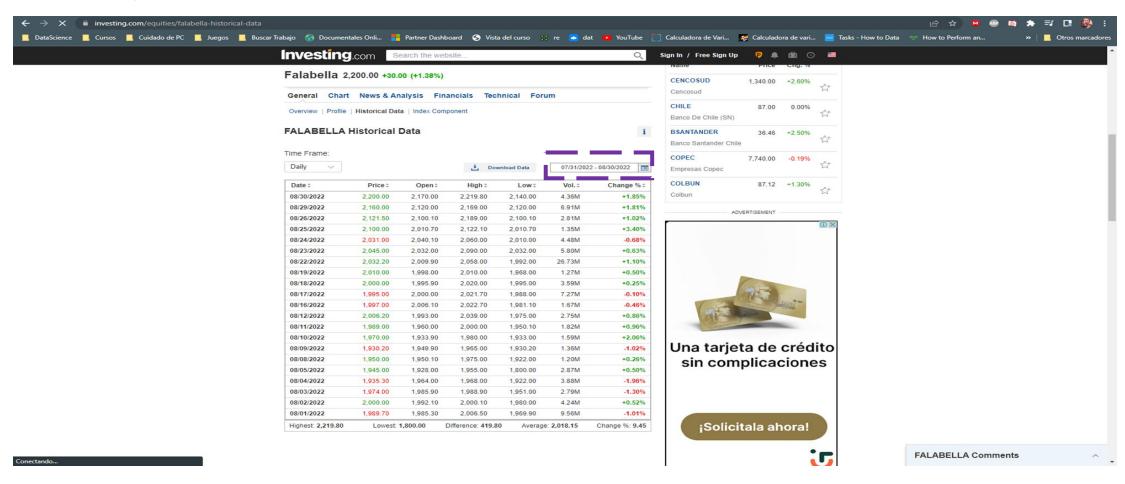
3. Hacemos un "AVPÁG" para bajar una pantalla y poder interaccionar con el link "Historical Data"



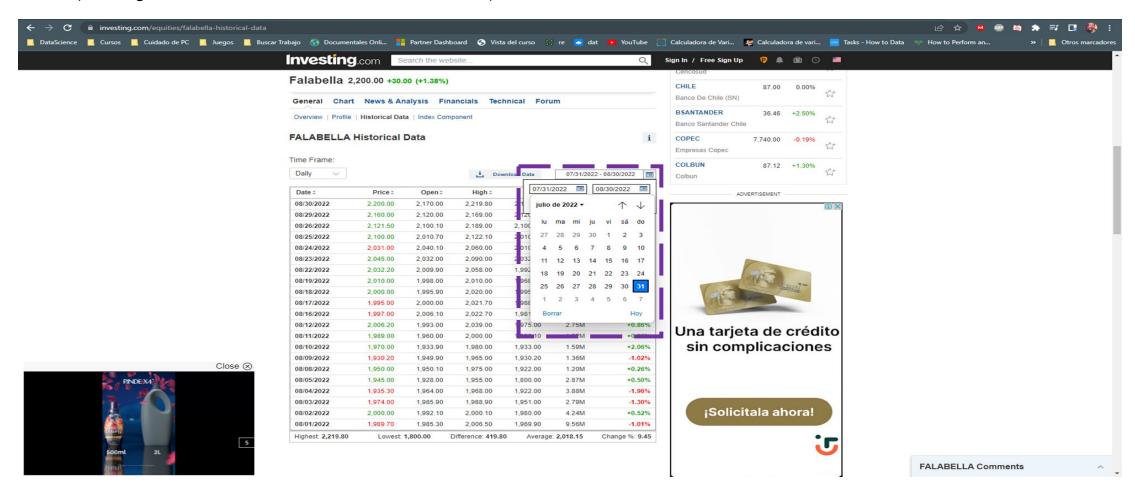
4. Hacemos un "AVPÁG" para bajar una pantalla y poder interaccionar con la data histórica.



Queremos interactuar con la barra de periodos, para poder solicitar toda la data del año 2022.
 (Investigando cómo interactuar con este elemento)



Queremos interactuar con la barra de periodos, para poder solicitar toda la data del año 2022.
 (Investigando cómo interactuar con este elemento)



6. Descargamos con BS4 la información de esta tabla y la almacenaremos en la carpeta de la Empresa.

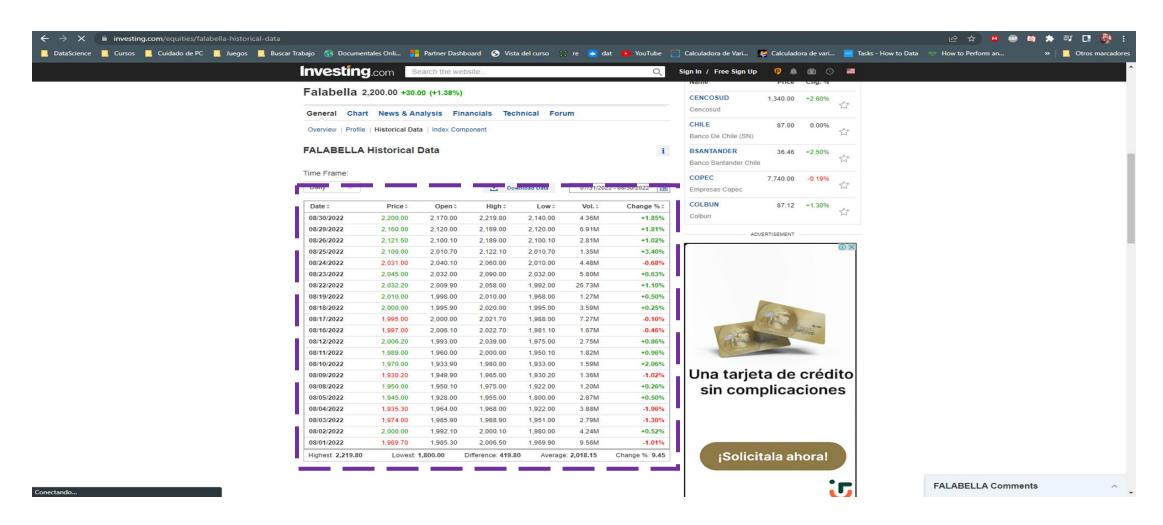
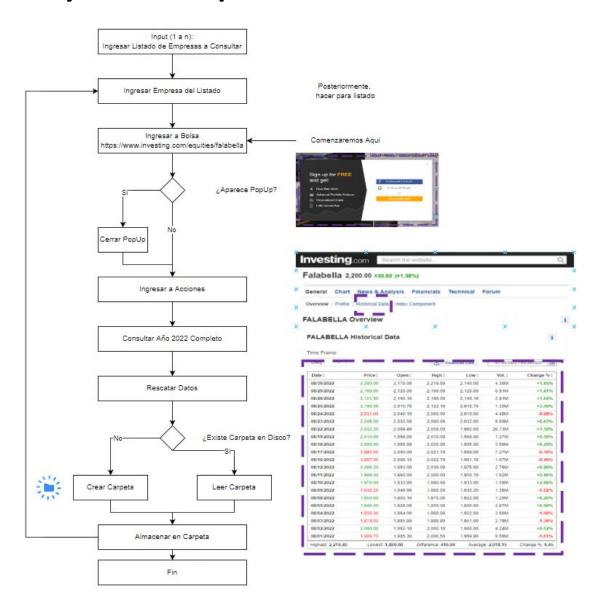


Diagrama del proceso / Pseudo código (Opción 2) Proyecto – Grupo 2



Estructura archivo / Output / Docs Proyecto – Grupo 2

EJ: Precio de la Acción de Falabella en el Tiempo:

1 A	В	C	D	Е	F	G
Date	Price	Open	High	Low	Vol	Change_Percent
30-08-2022	2200	2170	2219,8	2140	4.36M	1,85
29-08-2022	2160	2120	2169	2120	6.91M	1,81
26-08-2022	2121,5	2100,1	2189	2100,1	2.81M	1,02
25-08-2022	2100	2010,7	2122,1	2010,7	1.35M	3,4
24-08-2022	2031	2040,1	2060	2010	4.48M	-0,68
23-08-2022	2045	2032	2090	2032	5.80M	0,63
22-08-2022	2032,2	2009,9	2058	1992	26.73M	1,1
19-08-2022	2010	1998	2010	1968	1.27M	0,5
18-08-2022	2000	1995,9	2020	1995	3.59M	0,25
17-08-2022	1995	2000	2021,7	1988	7.27M	-0,1
16-08-2022	1997	2006,1	2022,7	1981,1	1.67M	-0,46
12-08-2022	2006,2	1993	2039	1975	2.75M	0,86
11-08-2022	1989	1960	2000	1950,1	1.82M	0,96
10-08-2022	1970	1933,9	1980	1933	1.59M	2,06
09-08-2022	1930,2	1949,9	1965	1930,2	1.36M	-1,02
08-08-2022	1950	1950,1	1975	1922	1.20M	0,26
05-08-2022	1945	1928	1955	1800	2.87M	0,5
04-08-2022	1935,3	1964	1968	1922	3.88M	-1,96
03-08-2022	1974	1985,9	1988,9	1951	2.79M	-1,3
02-08-2022	2000	1992,1	2000,1	1980	4.24M	0,52
01-08-2022	1989,7	1985,3	2006,5	1969,9	9.56M	-1,01

Nombre	Tipo	Scrap	
Date	Date	SI	
Price	Float	SI	
Open	Float	SI	
High	Float	SI	
Low	Float	SI	
Vol.	Str	NO	
Change_Percent	Float	SI	

Número de Filas: Inicialmente, ojala todo el 2022

Se genera una fila por dia, con 5/ columnas a rescatar por empresa

Carpeta / Docs Proyecto – Grupo 2



Pre-presentación

Objetivo: Justificar la elección de tecnología e infraestructura

Deadline: 1 septiembre 23:59

- Presentación con slides
- Explicar los pasos que se seguirán
- Mostrar un borrador del scraper en pseudocódigo
- Borrador del csv final (número de filas, y columnas)
- Base de la presentación final del proyecto

Pre-presentación

Obligatorio:

- Crear un producto nuevo
- Agregar el pdf de los slides en /docs
- Agregar un users.txt