

29/11/2023

Universidad de Guadalajara

Centro Universitario de Ciencias Exactas e Ingenierías

Ingeniería en Informática

Seguridad de la Información

Sandoval Chávez Miguel Ángel

Código: 220792825

Profesor: Guzmán Montes Carlos Alberto

Sección: D04

Ciclo escolar: 2023B

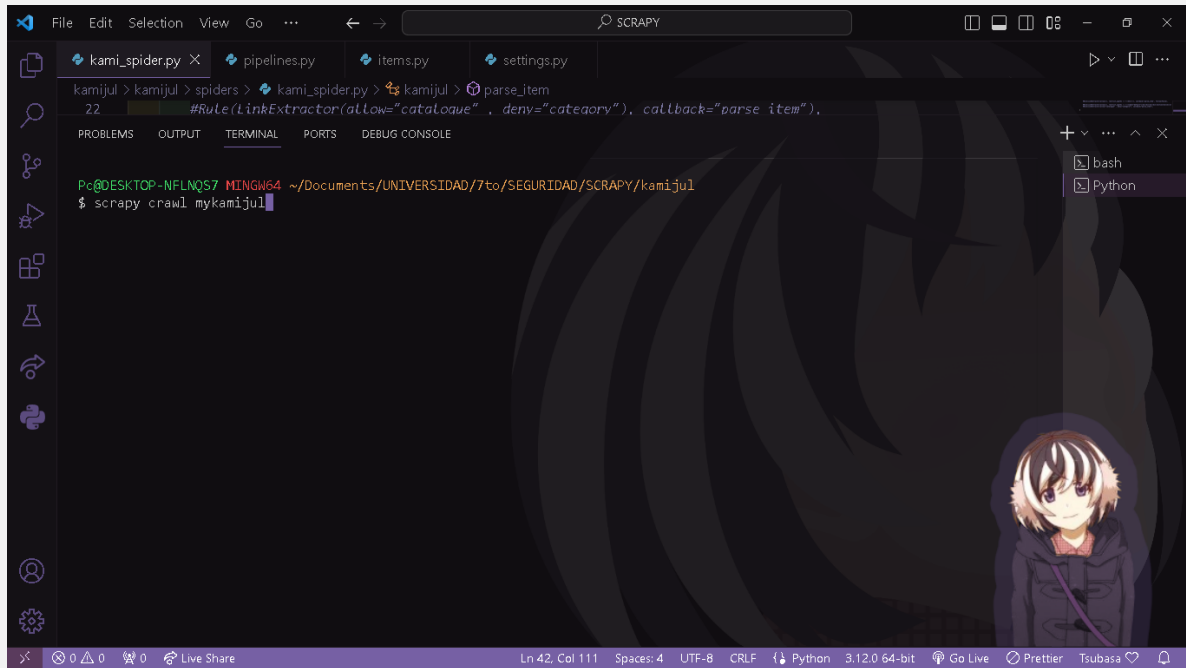
Contenido

Web Crawler con Scrapy	1
------------------------------	---

ES CONFIDENCIAL ONEE-SAN!!!

Web Crawler con Scrapy

Aquí se mostrará con capturas el funcionamiento del crawler utilizando scrapy, pandas y matplotlib.pyplot



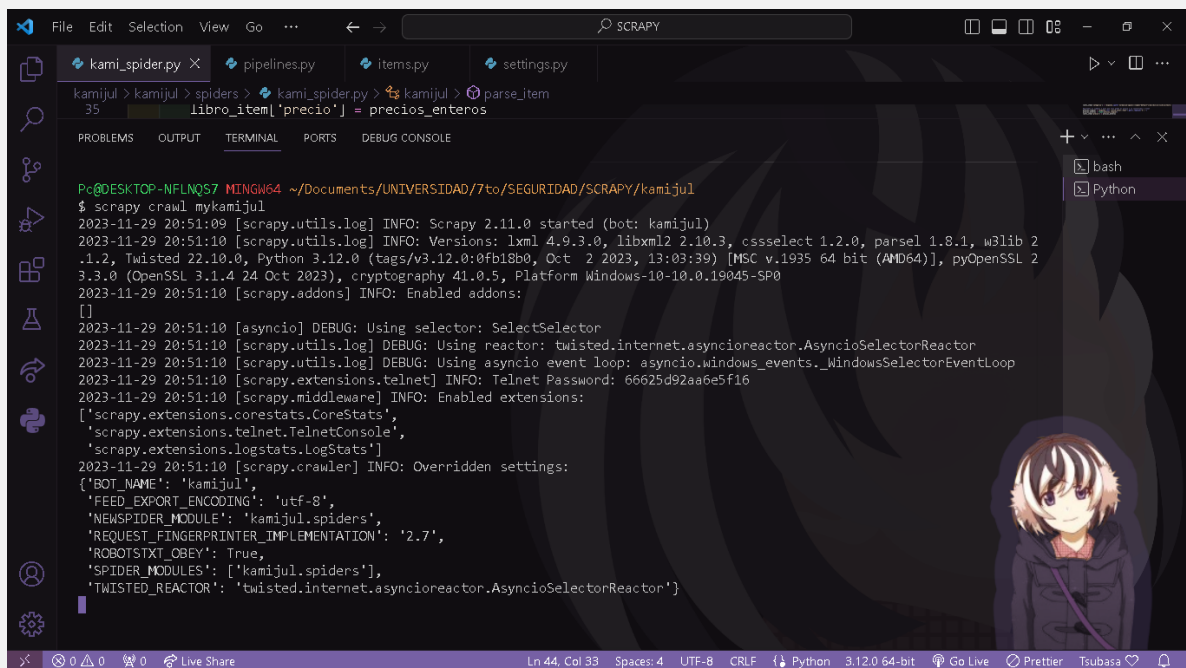
The screenshot shows a VS Code editor with a dark theme. The file explorer on the left shows a project named 'kamijul' with files 'kami_spider.py', 'pipelines.py', 'items.py', and 'settings.py'. The main editor window shows the 'kami_spider.py' file with the following code:

```
kamijul > kamijul > spiders > kami_spider.py > kamijul > parse_item
22 #Rule(LinkExtractor(allow="catalogue", deny="category"), callback="parse_item").
```

The terminal window at the bottom shows the command prompt and the execution of the scrapy crawl command:

```
Pc@DESKTOP-NFLNQS7 MINGW64 ~/Documents/UNIVERSIDAD/7to/SEGURIDAD/SCRAPY/kamijul
$ scrapy crawl mykamijul
```

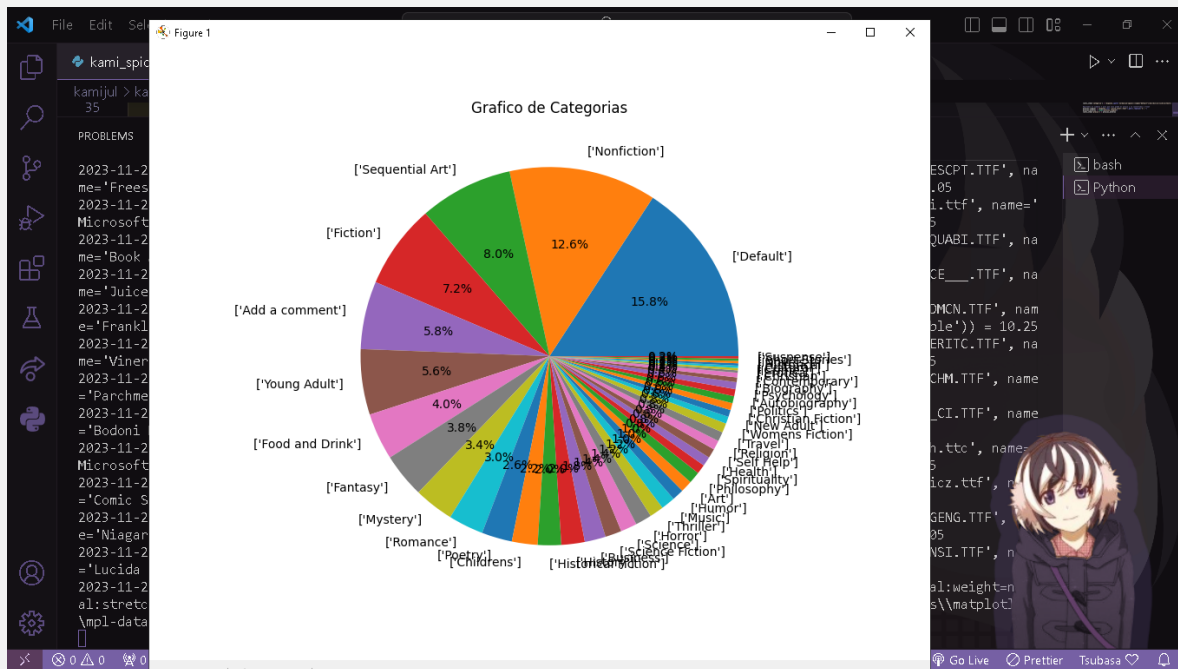
Comenzamos con el comando para iniciar el crawler (se debe estar en el directorio del proyecto)



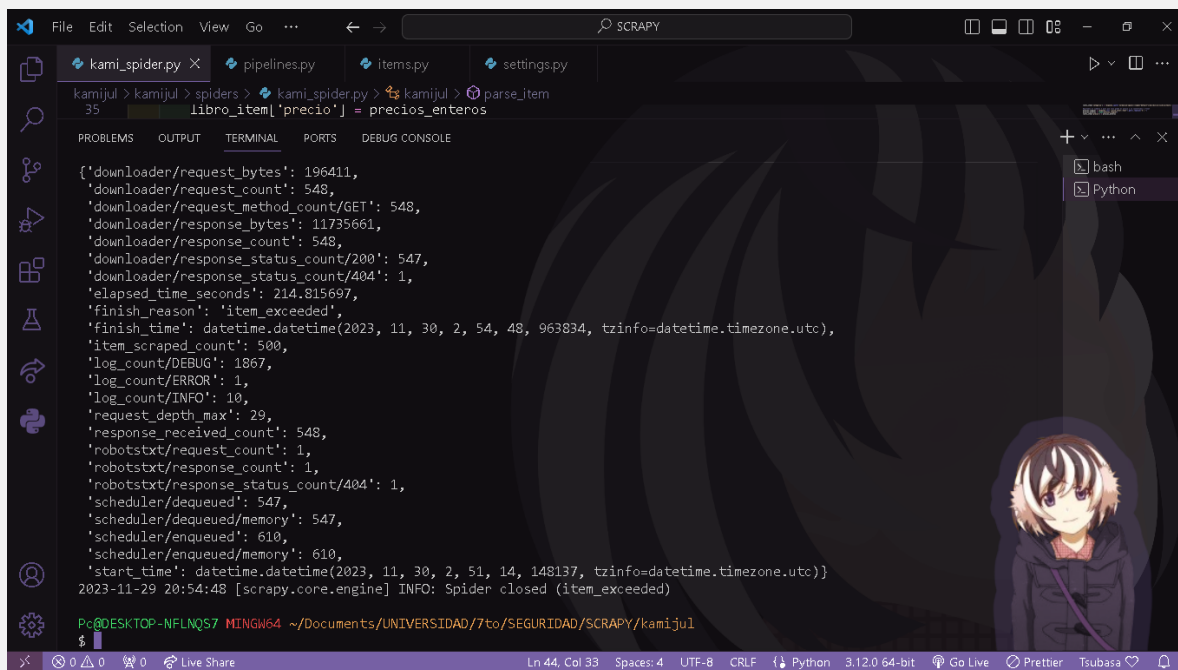
The screenshot shows the same VS Code editor as before, but the terminal window now displays the output of the scrapy crawl command:

```
Pc@DESKTOP-NFLNQS7 MINGW64 ~/Documents/UNIVERSIDAD/7to/SEGURIDAD/SCRAPY/kamijul
$ scrapy crawl mykamijul
2023-11-29 20:51:09 [scrapy.utils.log] INFO: Scrapy 2.11.0 started (bot: kamijul)
2023-11-29 20:51:10 [scrapy.utils.log] INFO: Versions: lxml 4.9.3.0, libxml2 2.10.3, cssselect 1.2.0, parsel 1.8.1, w3lib 2
.1.2, Twisted 22.10.0, Python 3.12.0 (tags/v3.12.0:0fb18b0, Oct 2 2023, 13:03:39) [MSC v.1935 64 bit (AMD64)], pyOpenSSL 2
3.3.0 (OpenSSL 3.1.4 24 Oct 2023), cryptography 41.0.5, Platform Windows-10-10.0.19045-SP0
2023-11-29 20:51:10 [scrapy.addons] INFO: Enabled addons:
[]
2023-11-29 20:51:10 [asyncio] DEBUG: Using selector: SelectSelector
2023-11-29 20:51:10 [scrapy.utils.log] DEBUG: Using reactor: twisted.internet.asyncioreactor.AsyncioSelectorReactor
2023-11-29 20:51:10 [scrapy.utils.log] DEBUG: Using asyncio event loop: asyncio.windows_events._WindowsSelectorEventLoop
2023-11-29 20:51:10 [scrapy.extensions.telnet] INFO: Telnet Password: 66625d92aa6e5f16
2023-11-29 20:51:10 [scrapy.middleware] INFO: Enabled extensions:
['scrapy.extensions.corestats.CoreStats',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.logstats.LogStats']
2023-11-29 20:51:10 [scrapy.crawler] INFO: Overridden settings:
{'BOT_NAME': 'kamijul',
'FEED_EXPORT_ENCODING': 'utf-8',
'NEWSPIDER_MODULE': 'kamijul.spiders',
'REQUEST_FINGERPRINTER_IMPLEMENTATION': '2.7',
'ROBOTSTXT_OBEY': True,
'SPIDER_MODULES': ['kamijul.spiders'],
'TWISTED_REACTOR': 'twisted.internet.asyncioreactor.AsyncioSelectorReactor'}
```

Esperamos unos momentos.



Después nos mostrará esta ventana que es una gráfica de la columna 'categoría' en la que se muestran en porcentaje. La podemos cerrar por detrás ya se creó una imagen .png de esta misma.



```

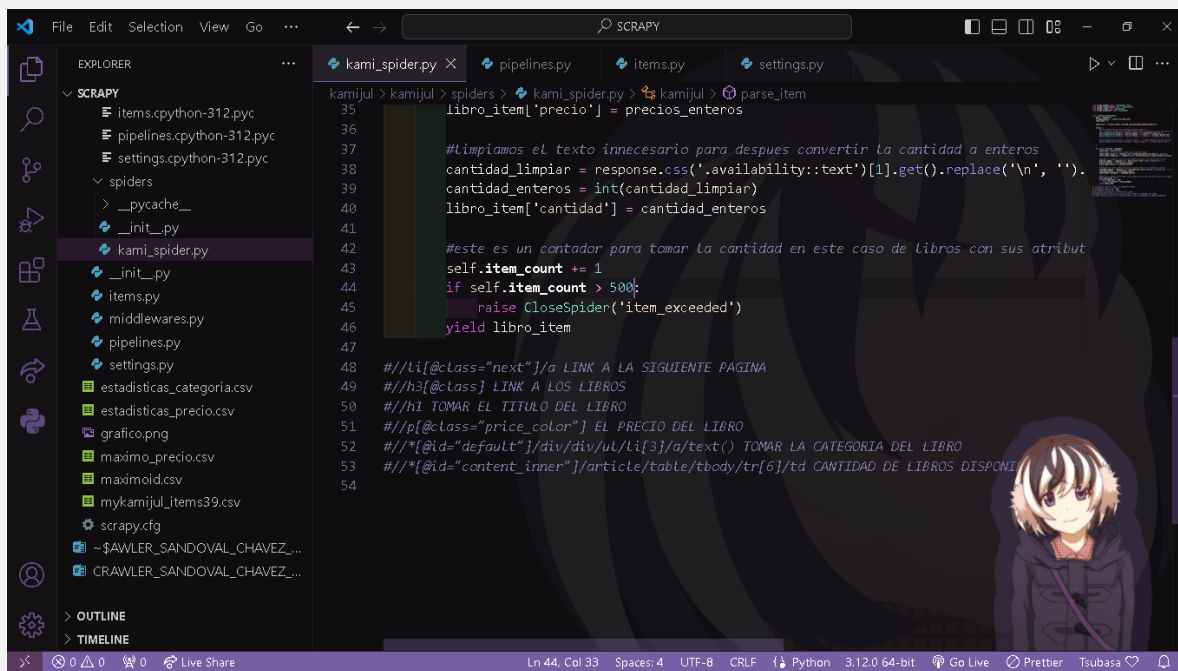
File Edit Selection View Go ... SCRAPY
kami_spider.py X pipelines.py items.py settings.py
kami > kami > spiders > kami_spider.py > kami > parse_item
35 libro_item['precio'] = precios_enteros

PROBLEMS OUTPUT TERMINAL PORTS DEBUG CONSOLE
[{'downloader/request_bytes': 196411,
'downloader/request_count': 548,
'downloader/request_method_count/GET': 548,
'downloader/response_bytes': 11735661,
'downloader/response_count': 548,
'downloader/response_status_count/200': 547,
'downloader/response_status_count/404': 1,
'elapsed_time_seconds': 214.815697,
'finish_reason': 'item_exceeded',
'finish_time': datetime.datetime(2023, 11, 30, 2, 54, 48, 963834, tzinfo=datetime.timezone.utc),
'item_scraped_count': 500,
'log_count/DEBUG': 1867,
'log_count/ERROR': 1,
'log_count/INFO': 10,
'request_depth_max': 29,
'response_received_count': 548,
'robotstxt/request_count': 1,
'robotstxt/response_count': 1,
'robotstxt/response_status_count/404': 1,
'scheduler/dequeued': 547,
'scheduler/dequeued/memory': 547,
'scheduler/enqueued': 610,
'scheduler/enqueued/memory': 610,
'start_time': datetime.datetime(2023, 11, 30, 2, 51, 14, 148137, tzinfo=datetime.timezone.utc)}]
2023-11-29 20:54:48 [scrapy.core.engine] INFO: Spider closed (item_exceeded)

Pc@DESKTOP-NFLNQS7 MINGW64 ~/Documents/UNIVERSIDAD/7to/SEGURIDAD/SCRAPY/kamijul
$

```

Aquí se mira que el crawler ya terminó.



```

File Edit Selection View Go ... SCRAPY
EXPLORER ... kami_spider.py X pipelines.py items.py settings.py
SCRAPY
  items.python-312.pyc
  pipelines.python-312.pyc
  settings.python-312.pyc
  spiders
    > __pycache__
    > __init__.py
    > kami_spider.py
    > __init__.py
    > items.py
    > middlewares.py
    > pipelines.py
    > settings.py
  estadisticas_categoria.csv
  estadisticas_precio.csv
  grafico.png
  maximo_precio.csv
  maxmold.csv
  mykamiul_items39.csv
  scrapy.cfg
  ~$AWLER_SANDOVAL_CHAVEZ...
  CRAWLER_SANDOVAL_CHAVEZ...

  > OUTLINE
  > TIMELINE

kami > kami > spiders > kami_spider.py > kami > parse_item
35 libro_item['precio'] = precios_enteros
36
37
38 #limpiamos el texto innecesario para despues convertir la cantidad a enteros
39 cantidad_limpiar = response.css('.availability::text')[1].get().replace('\n', '')
40 cantidad_enteros = int(cantidad_limpiar)
41 libro_item['cantidad'] = cantidad_enteros
42
43 #este es un contador para tomar la cantidad en este caso de libros con sus atribut
44 self.item_count += 1
45 if self.item_count > 500:
46     raise CloseSpider('item_exceeded')
47 yield libro_item
48
49 #//li[@class="next"]/a LINK A LA SIGUIENTE PAGINA
50 #//h3[@class] LINK A LOS LIBROS
51 #//h1 TOMAR EL TITULO DEL LIBRO
52 #//p[@class="price_color"] EL PRECIO DEL LIBRO
53 #//*(@id="default")/div/div/ul/li[3]/a/text() TOMAR LA CATEGORIA DEL LIBRO
54 #//*(@id="content_inner")/article/table/tbody/tr[6]/td CANTIDAD DE LIBROS DISPONI

```

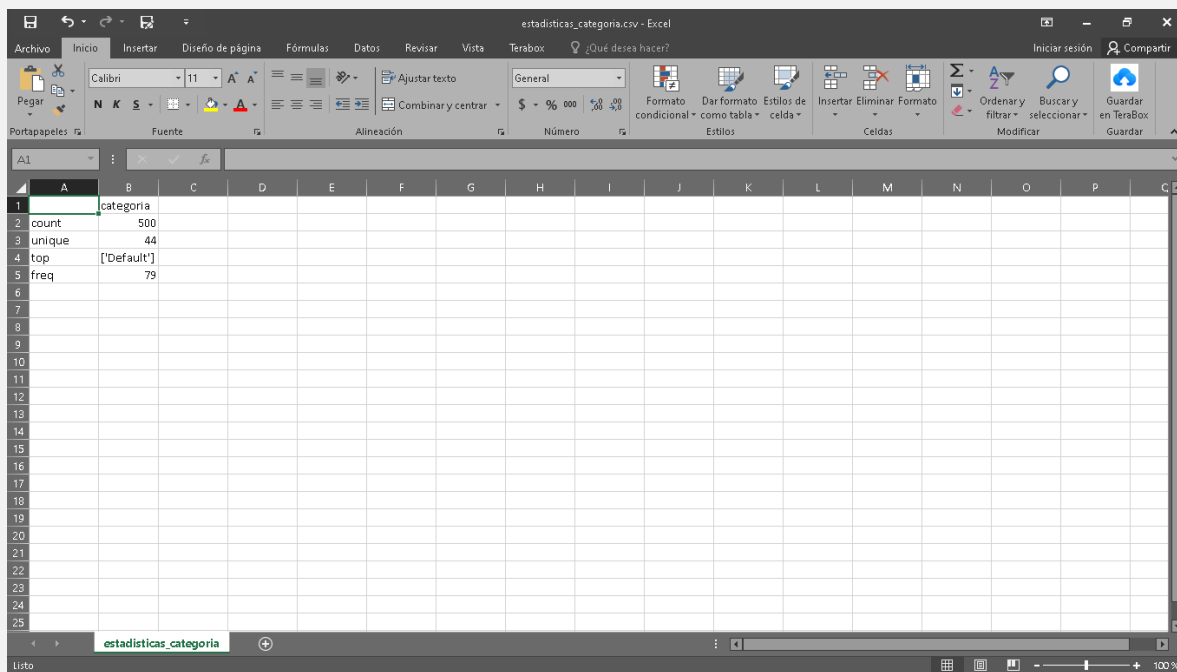
en la parte izquierda se puede apreciar que se craron varios archivos .csv con la

información que el crawler se trajo de la página

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
	titulo	precio	categoria	cantidad												
1	titulo	precio	categoria	cantidad												
2	The Coming	17.93	Default	19												
3	A Light in the	51.77	Poetry	22												
4	Tipping the \	53.74	Historical Fic	20												
5	Soumission	50.1	Fiction	20												
6	The Boys in t	22.6	Default	19												
7	Sapiens: A Bi	54.23	History	20												
8	The Black Me	52.15	Poetry	19												
9	The Requien	22.65	Young Adult	19												
10	The Dirty Litt	33.34	Business	19												
11	Sharp Object	47.82	Mystery	20												
12	Shakespeare	20.66	Poetry	19												
13	Scott Pilgrim	52.29	Sequential A	19												
14	Rip it Up and	35.02	Music	19												
15	Starving Hea	13.99	Default	19												
16	It's Only the	45.17	Travel	19												
17	Set Me Free	17.46	Young Adult	19												
18	Libertarianis	51.33	Politics	19												
19	Behind Close	52.22	Thriller	18												
20	In Her Wake	12.84	Thriller	19												
21	You can't bur	33.63	Poetry	17												
22	In a Dark, Da	19.63	Mystery	18												
23	Maude (1885	18.02	Default	18												
24	Penny Mayb	33.29	Default	18												
25	Sophie's Wo	15.94	Philosophy	18												

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
483	Silence in th	58.33	Suspense	8												
484	Being Mortal	55.06	Nonfiction	8												
485	The Great Ga	36.05	Default	7												
486	Shadows of t	39.67	Christian Fic	8												
487	Kierkegaard:	47.13	Philosophy	8												
488	A Murder Ov	13.2	Nonfiction	8												
489	The Good Gi	49.03	Default	7												
490	John Vassos:	20.22	Default	8												
491	Nightstruck:	50.35	Young Adult	7												
492	The Glass Ca	16.24	Add a comm	7												
493	32 Yolks	53.63	Food and Dri	8												
494	Naturally Lea	11.38	Food and Dri	7												
495	The Faith of	39.55	Biography	7												
496	"Most Blesse	44.48	History	8												
497	I'll Give You	56.48	Default	8												
498	I Will Find Yc	44.21	Nonfiction	8												
499	You Are a Ba	12.08	Self Help	7												
500	The Drownin	35.67	Add a comm	7												
501	Hystopia: A N	21.96	Fiction	8												
502																
503																
504																
505																
506																
507																

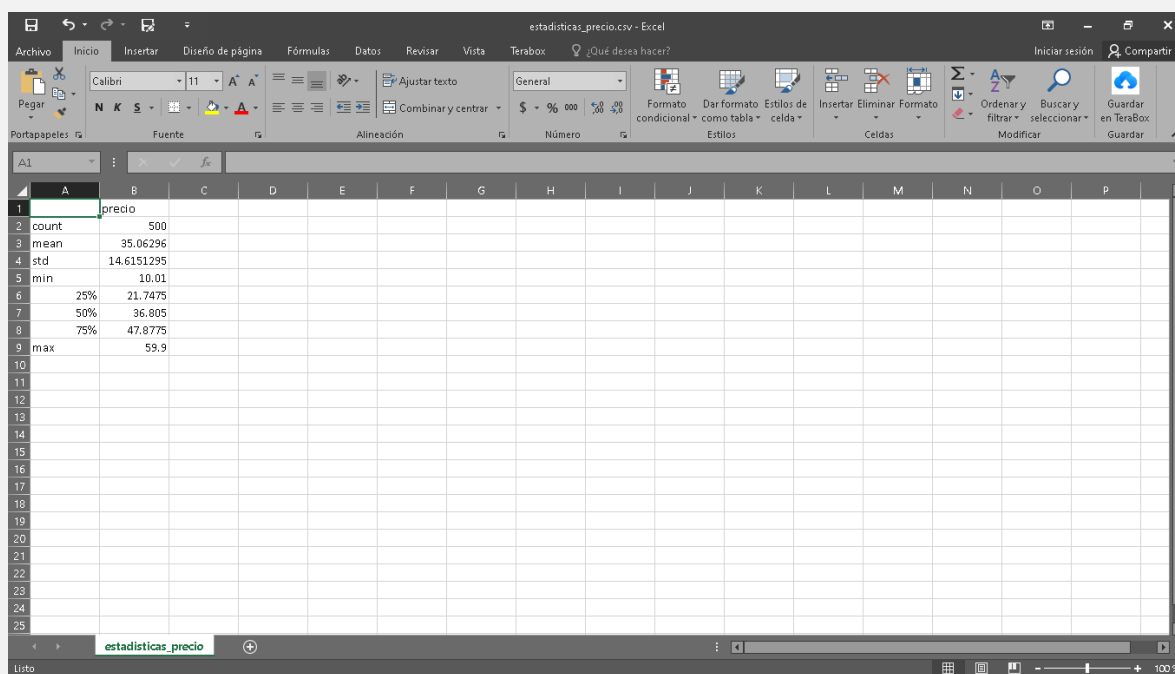
Se muestran en el archivo principal que se trajo 500 en este caso libros.



estadisticas_categoria.csv - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		categoria															
2	count	500															
3	unique	44															
4	top	['Default']															
5	freq	79															
6																	
7																	
8																	
9																	
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	

estadisticas_categoria



estadisticas_precio.csv - Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		precio															
2	count	500															
3	mean	35.06296															
4	std	14.6151295															
5	min	10.01															
6	25%	21.7475															
7	50%	36.805															
8	75%	47.8775															
9	max	59.9															
10																	
11																	
12																	
13																	
14																	
15																	
16																	
17																	
18																	
19																	
20																	
21																	
22																	
23																	
24																	
25																	

estadisticas_precio

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	360	[The Diary of a Young Girl]	[Nonfiction]	59.9	12										
2	[The Diary of a Young Girl]	[Nonfiction]	59.9	12											
3	[Nonfiction]	59.9	12												
4	59.9	12													
5	12														
6															
7															
8															
9															
10															
11															
12															
13															
14															
15															
16															
17															
18															
19															
20															
21															
22															
23															
24															
25															

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	1	[A Light in the Attic]	[Poetry]	51.77	22											
2	[A Light in the Attic]	[Poetry]	51.77	22												
3	[Poetry]	51.77	22													
4	51.77	22														
5	22															
6																
7																
8																
9																
10																
11																
12																
13																
14																
15																
16																
17																
18																
19																
20																
21																
22																
23																
24																
25																

Aquí se muestran los otros archivos que se derivaron del principal.