



Scrapy

<https://scrapy.org/>

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/

User-agent: Applebot
Disallow: /ajax/
Disallow: /album.php
Disallow: /checkpoint/
Disallow: /contact_importer/
Disallow: /feeds/
Disallow: /file_download.php
Disallow: /hashtag/
Disallow: /l.php
Disallow: /live/
Disallow: /moments_app/
Disallow: /p.php
Disallow: /photo.php
Disallow: /photos.php
Disallow: /sharer/

Robots.txt

Allow All

User-agent: *

Disallow:

Disallow All

User-agent: *

Disallow: /

Specific Directions

User-agent: *

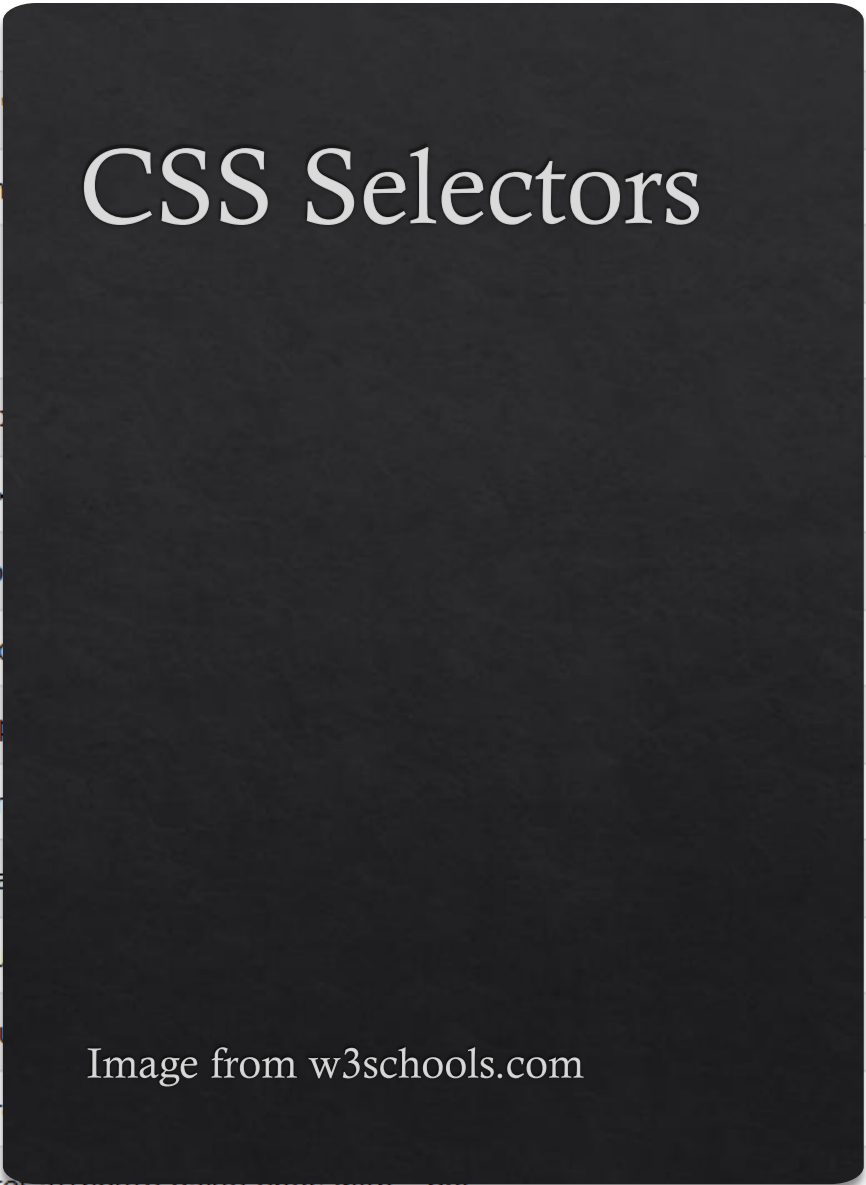
Disallow: /admin/

Disallow: /user/file.html

Crawl-delay: 10

Use our [CSS Selector Tester](#) to demonstrate the different selectors.

Selector	Example	Example description
<u>.class</u>	.intro	Selects all elements with class="intro"
<u>#id</u>	#firstname	Selects the element with id="firstname"
<u>*</u>	*	Selects all elements
<u>element</u>	p	Selects all <p> elements
<u>element,element</u>	div, p	Selects all <div> elements and all <p> elements
<u>element element</u>	div p	Selects all <p> elements inside <div> elements
<u>element>element</u>	div > p	Selects all <p> elements where the p element is the first child of the div element
<u>element+element</u>	div + p	Selects all <p> elements that are placed directly after the <div> element
<u>element1~element2</u>	p ~ ul	Selects every element that are placed after the <p> element
<u>[attribute]</u>	[target]	Selects all elements with a target attribute
<u>[attribute=value]</u>	[target=_blank]	Selects all elements with target="_blank"
<u>[attribute~=value]</u>	[title~=flower]	Selects all elements with a title attribute value starting with "flower"
<u>[attribute =value]</u>	[lang =en]	Selects all elements with a lang attribute value starting with "en"
<u>[attribute^=value]</u>	a[href^="https"]	Selects every <a> element whose href attribute value starts with "https"
<u>[attribute\$=value]</u>	a[href\$=".pdf"]	Selects every <a> element whose href attribute value ends with ".pdf"
<u>[attribute*=value]</u>	a[href*="w3schools"]	Selects every <a> element whose href attribute value contains the substring "w3schools"



```
<html>
```

```
<style>
```

```
span {color:red;}
```

```
</style>
```

```
<span>Get me</span>
```

```
<span style="color:green;">Get me</span>
```

```
</html>
```

::text Pseudo Pseudo selector

```
In [5]: Selector(text='<span>get me</span>').css('span').get()
```

```
Out[5]: '<span>get me</span>'
```

```
In [6]: Selector(text='<span>get me</span>').css('span::text').get()
```

```
Out[6]: 'get me'
```


Tonight's Project



SEARCH RESULTS

CONNECT WITH US:

MY HOME FILE



PROPERTY SEARCH

BUYING

SELLING

RENTING

RELOCATING

REVISE

SAVE

NEW

152 properties found: Harford County, MD; \$300,000-\$400,000; 1+ Bedrooms;



\$399,999



Scenic Manor
515 Dusk View Drive, Havre De Grace
4 Beds, 2.1 Baths

[» More Info](#)



\$399,999



Gunpowder
1609 Bridewells Court, Joppa
4 Beds, 3.1 Baths

[» More Info](#)

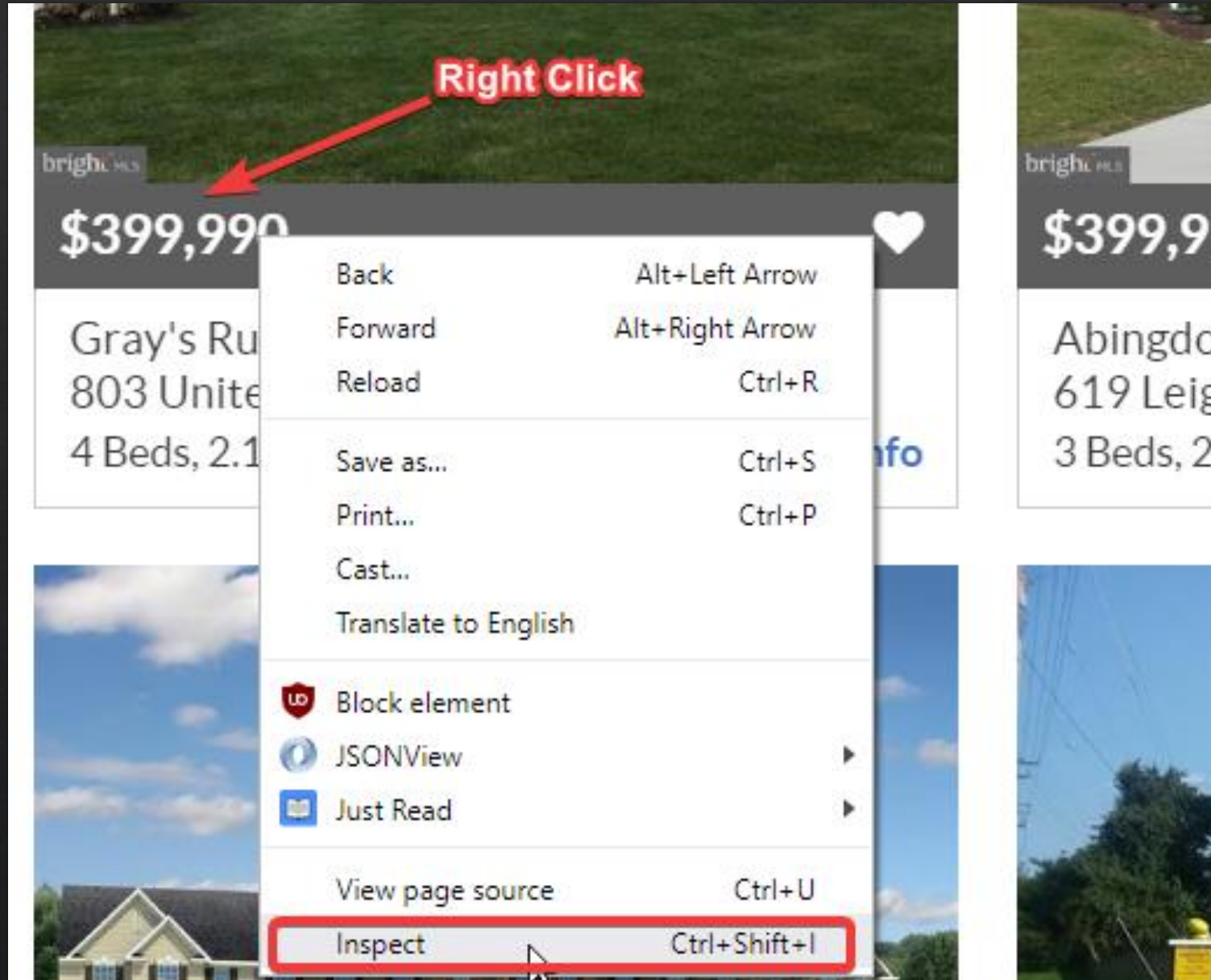


\$399,990

Advocate Hill Farm
1616 Cynthia Court, Joppa
3 Beds, 2.1 Baths



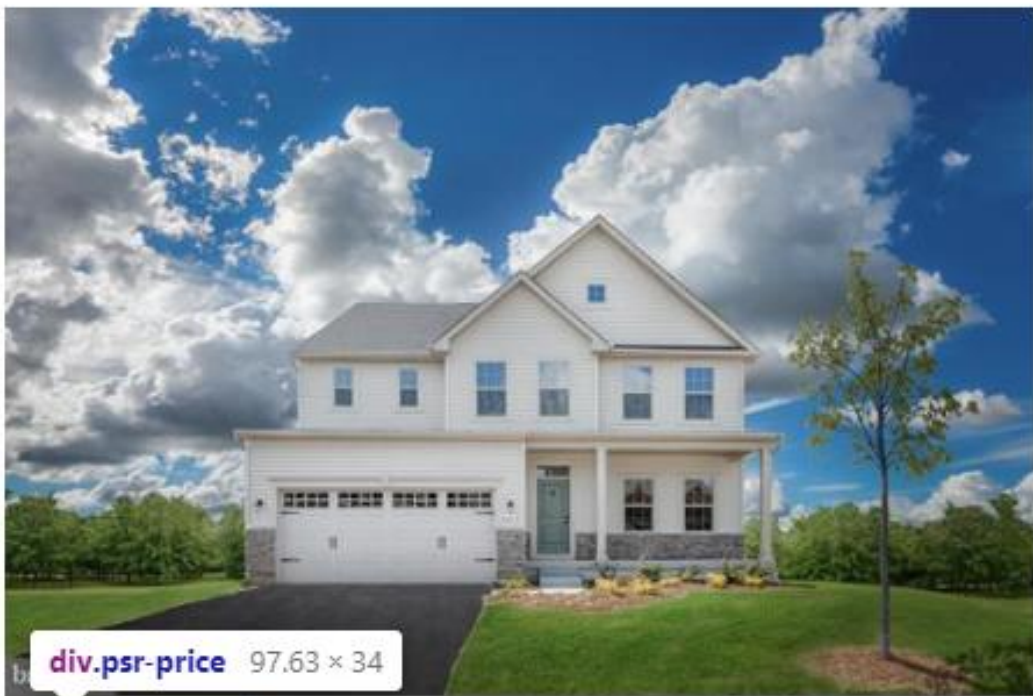
Inspect Element



Price Class: psr-price

Sort By: Price (high to low)

Properties found: Harford County, MD; \$300,000-\$400,000; 1+ Bedrooms;



div.psr-price 97.63 × 34

\$399,999

Scenic Manor

515 Dusk View Drive, Havre De Grace

4 Beds, 2.1 Baths

[» More Info](#)

[illegible]

Address

bright MLS

\$200,000

span.psr-address 424 × 46


Scenic Manor
515 Dusk View Drive, Havre De Grace

4 Beds, 2.1 Baths

[» More Info](#)

```
▼ <span class="psr-address"> == $0
  <span>Scenic Manor</span>
  <br>
  <span>515 Dusk View Drive, Havre De Grace</span>
</span>
▶ <div class="psr-more clearfix">...</div>
  ::after
</div>
</div>
▶ <div class="columns psr-result">...</div>
```

Details



brigh MLS

1 / 6

#text 40.25 × 19

Beds/Baths:	3/2.1	MLS#:	MDHR201902
Interior Sq. Ft:	1,680	Subdivision:	Abingdon
Acreage:	0.25	Design/Type:	Detached
Age:	0 years		
Style:	Bi-level		

[← BACK TO RESULTS](#)

[MAKE APPOINTMENT](#)

[PROPERTY INFO](#) [LOCATION/MAP](#) [SCHOOLS](#) [COMMUNITY](#)

```
<h2 class="property-location clearfix">...</h2>
<div role="region" aria-label="Slideshow of property for sale">...</div>
<div class="tour clearfix">...</div>
<div class="primary-details clearfix">
  ::before
  <div class="primary-col-btns clearfix view-not-print">...</div>
  <div class="primary-col1"> == $0
    <strong>Beds/Baths:</strong>
    " 3/2.1"
    <br>
    <strong>Interior Sq. Ft:</strong>
    " 1,680"
    <br>
    <strong>Acreage:</strong>
    " 0.25"
    <br>
    <strong>Age:</strong>
    "0 "
    <span>years</span>
    <br>
    <strong>Style:</strong>
    <br class="br-oo">
    "Bi-level  "
  </div>
  <div class="primary-col2">...</div>
  ::after
</div>
<div class="secondary-details" id="tabsPD">...</div>
<div class="row">...</div>
<div class="row contact-details">...</div>
<section class="broker-reciprocity">...</section>
</div>
```


Code Demo

Spider Boilerplate

```
1  import scrapy
2
3  class MySpiderSpider(scrapy.Spider):
4      name = 'MySpider'
5      allowed_domains = ['obscurefjord.com']
6      start_urls = ['http://obscurefjord.com/']
7
8      def parse(self, response):
9          pass
10
```

Single Page

```
1  # -*- coding: utf-8 -*-
2  import scrapy
3
4
5  class HomelistSpider(scrapy.Spider):
6      name = 'homelist'
7      allowed_domains = ['https://pattersonschwartz.com']
8      start_urls = ['http://www.pattersonschwartz.com/forsale/Harford/priceMin_250000']
9
10     def parse(self, response):
11         for r in response.css('.psr-result'):
12             yield {
13                 'price': r.css('.psr-price::text').get(),
14                 'cdp': r.css('.psr-address > span:nth-of-type(1)::text').get(),
15                 'address': r.css('.psr-address > span:nth-of-type(2)::text').get(),
16                 'listingurl': r.css('.psr-more-info::attr(href)').get()
17             }
```


Multiple Pages

```
class HomelistapiSpider(scrapy.Spider):
    name = 'homelistapi'
    allowed_domains = ['pattersonschwartz.com']
    page = 1
    api_url = 'http://www.pattersonschwartz.com/api/ps/forsale/Cecil,Harford/priceMin_225000/priceMax_600000/page_{}'
    start_urls = [api_url.format(page)]

    def parse(self, response):
        data = json.loads(response.text)
        for r in data['thumbs']:
            yield{
                'id' : r['id'],
                'address' : r['a'],
                'price' : r['p'],
                'cdp' : r['c'],
                'picture' : r['pu']
            }
        if not data['isLastPage']:
            self.page += 1
            yield scrapy.Request(url=self.api_url.format(self.page), callback=self.parse)
```

Deprecated Slides

[← BACK TO RESULTS](#)[MAKE APPOINTMENT](#)

Beds/Baths: 6/4.2

Interior Sq. Ft: 10,700

Acreage: 2.34

Age: 90 years

Style: Tudor

MLS#: 1000377016

Subdivision: South Wayne


Design/Type: Detached

PROPERTY INFO

LOCATION/MAP

SCHOOLS

COMMUNITY

 Description

[← BACK TO RESULTS](#)[MAKE APPOINTMENT](#)

Beds/Baths: 0/0.0

Interior Sq. Ft: 10,700

Acreage: 1.03

Age: 2019 years

Zoning: R2

MLS#: MDHR222260


Subdivision: West Riding Farms

PROPERTY INFO

LOCATION/MAP

SCHOOLS

COMMUNITY

 Description

Beautiful flat 1.03 acre lot surrounded by Pine trees..Private yet close to everything!