

## Redfin 爬虫存在的主要问题及解释

“Redfin 资料需求.xlsx”表格中的数据可以分为两类，分别是**静态内容**和**动态内容**。静态内容是网页中直接包含在 HTML 文件中的内容，无需额外的用户交互或后续的脚本执行即可显示，页面加载后这些内容是固定的，可以直接解析。表格中以下内容为静态内容：

● FOR SALE - ACTIVE

3434 E Woodbine Rd, Orange, CA 92867

\$2,949,900

Est. \$19,557/mo

5

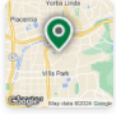
Beds

4.5

Baths

4,782

Sq Ft



For Sale 框内的 Beds, Baths, Sq Ft, Est price.

### About this home

Presented by **Maryam Amiri** | **REDFIN**

“A Luxurious Haven with Breathtaking Views nestled in the prestigious gated community of Hillcrest Estates, this stunning manor is a masterpiece of design and craftsmanship, featuring panoramic vistas from the rugged mountains to the shimmering ocean. Freshly painted with...”

[Show more](#) ▾

🕒

13 days on Redfin

🏠

Single-family

🔧

Built in 1984

📏

10,000 sq ft lot

💰

\$617 per sq ft

💵

\$375 monthly HOA fee

🚗

3 garage spaces (6 total)

🌡️

Has A/C

🧺

Washer and dryer hookups

👤

2% buyer's agent fee

🏡

Hillcrest (HLCR)

Listed by **Maryam Amiri** • DRE #01804754 • Redfin

Listed by **Christopher Bistolas** • DRE #01838313 • Redfin

Redfin checked: **3 minutes ago** (Aug 29, 2024 at 10:16am)

• Source: CRMLS #OC24169523

About this home 的内容

反之，动态内容是指需要通过 JavaScript 执行、用户交互（如点击或滚动），或异步请求（如 AJAX）从服务器获取并显示的内容。页面初次加载时，这些内容通常不直接包含在 HTML 中，而是由客户端（浏览器）在页面加载后生成或请求。表格中以下内容为动态内容：

Sale and tax history for 3434 E Woodbine Rd		
Sale History	Tax History	
Today		
Aug 16, 2024	Listing Removed	—
Date	CRMLS #OC24151594	Price
Jul 25, 2024	Listed (Active)	*
Date	CRMLS #OC24151594	Price
Aug, 2024		
Aug 16, 2024	Listed (Active)	\$2,949,900
Date	CRMLS #OC24169523	Price

### Sale and tax history

## Property details for 3434 E Woodbine Rd

**Parking**

Uncovered Spaces: 3 nullAttached Garage...

---

**Interior**

null<a href='https://my.matterport.com/show/?m=B9ePvKPxNc7&mls=1' target='\_bl...

---

**Exterior**

Structure Type: House Roof: Flat Tile...

---

**Financial**

Assessments: Unknown...

---

**Utilities**

Sewer: Public Sewer Water Source: District/Public...

---

**Location**

### Property details

如果用户不进行交互，如不点击“tax history”或不展开 Exterior 的模块，那么里面的内容就不会显示，这就要求爬虫具备网页交互的能力。

Redfin 的爬虫协议支持获取以上的**静态内容**，但是会严格检测用户端浏览器的交互行为，如果用户存在快速交互和频繁开关浏览器等可疑行为，就会违反协议并被禁止访问所有网页的内容。

以下是我的爬虫遭到拦截返回的截图，实测间隔在 1-5 分钟的访问和交互均会被认定为恶意行为。

REDFIN.



There seems to be an issue...

Our usage behavior algorithm thinks you might be a robot.

Ensure you are accessing Redfin.com according to our [terms of usage](#).

Tips:

- If you are using a VPN, pause it while browsing Redfin.com.
- Make sure you have Javascript enabled in your web browser.
- Ensure your browser is up to date.

Please email [techsupport@redfin.com](mailto:techsupport@redfin.com) for help and include the following information:

```
Timestamp: 1724950811
Client IP: 47.149.54.215
User Agent: Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like
Gecko) HeadlessChrome/128.0.0.0 Safari/537.36
```

因此，如果要实现多个地址的数据爬取并规避上述检测行为，爬虫有以下要求：

1. 仅爬取静态页面的数据
2. 设置爬取间隔，也就是等待 5 分钟以上再进行下一个地址数据的爬取，或者在同一个页面下，每隔 5 分钟以上进行一次交互（如展开下拉目录）。