The Web Scraping Process

Clarke Bishop

BIG DATA ENGINEER

@ClarkeBishop www.clarkebishop.com

Overview

Human browsing versus web scraping

HTTP overview

URL hacking

Human Browsing Versus Web Scraping

Human Versus Web Scraper

Human/Browser	Web Scraper

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start_url

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	**Parse HTML**

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	Parse HTML
Review for Useful Information	**Extract Useful Information**

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	Parse HTML
Review for Useful Information	Extract Useful Information
Interpret	Transform or Aggregate

Human Versus Web Scraper

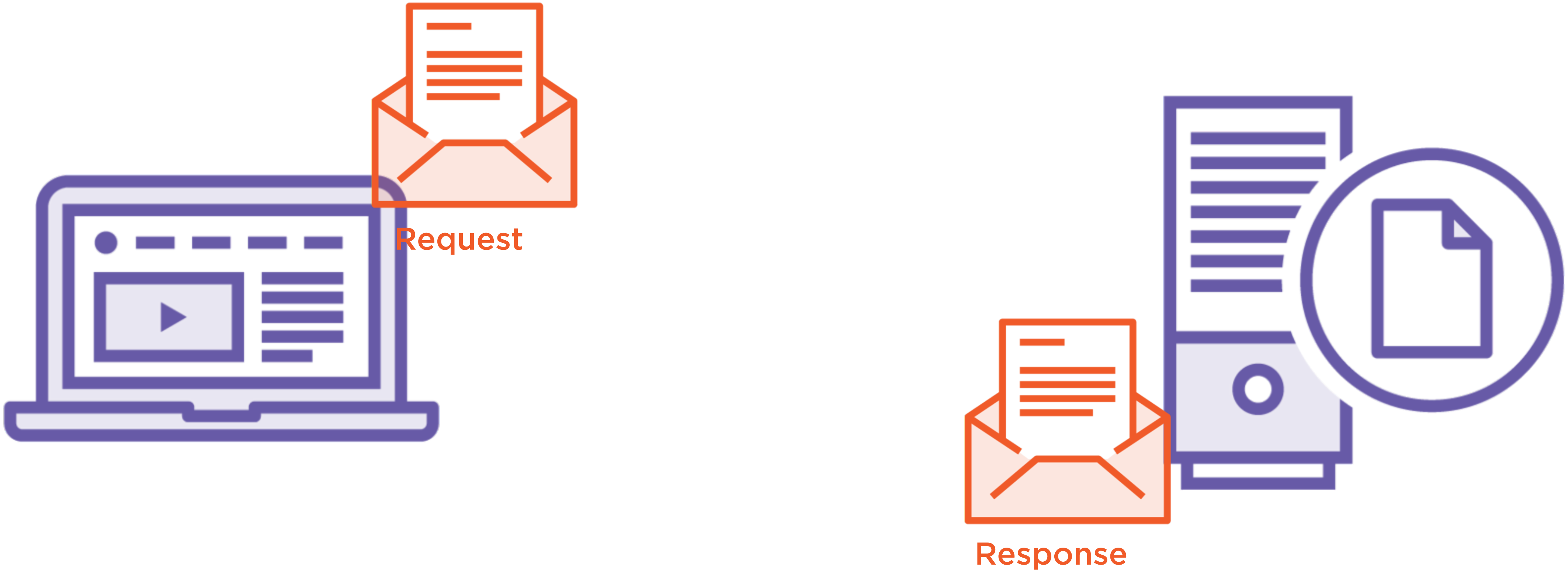
Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	Parse HTML
Review for Useful Information	Extract Useful Information
Interpret	Transform or Aggregate
Remember the Information	**Save the Data**

Human Versus Web Scraper

Human/Browser	Web Scraper
Enter a URL or click a bookmark	Set a start URL
Download HTML	Download HTML
Parse HTML & render	Parse HTML
Review for Useful Information	Extract Useful Information
Interpret	Transform or Aggregate
Remember the Information	Save the Data
Click a link-Enter another URL	Go the the next URL

HTTP Overview

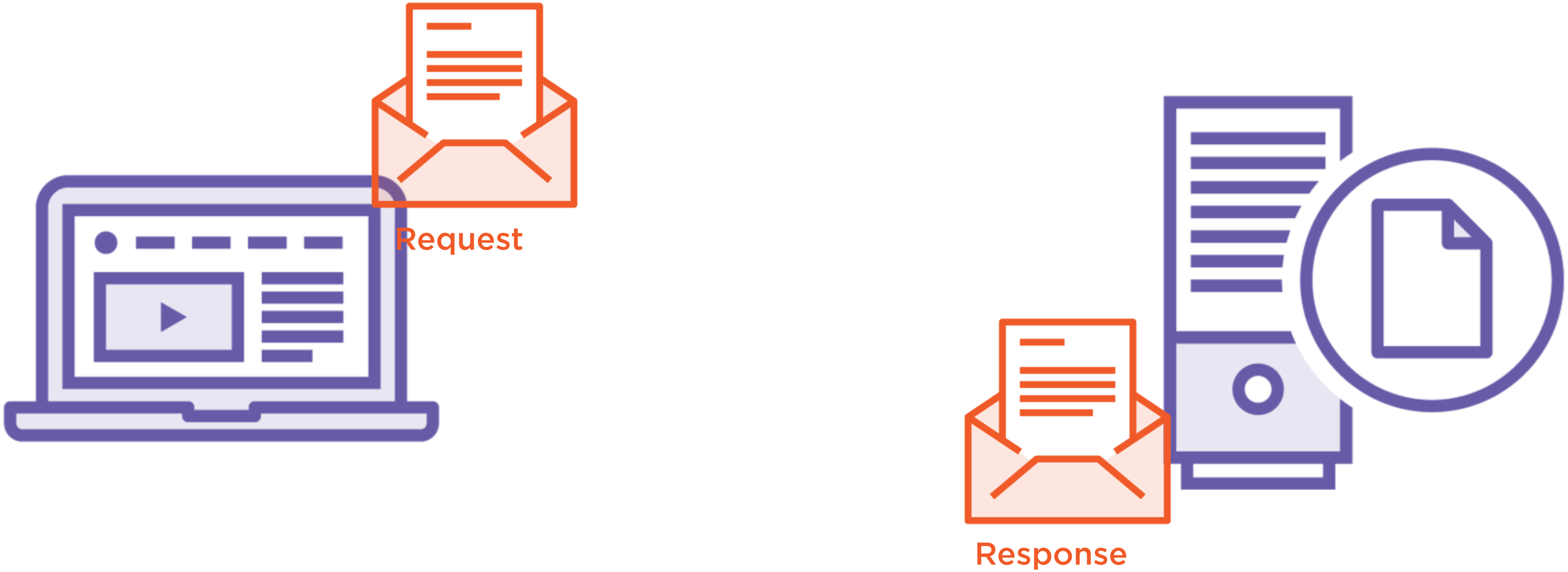
Request - Response



Request

Response

Request - Response



Hyper-Text Transfer Protocol (HTTP) is the protocol that powers the web.

HTTP Request

HTTP requests may include:

1 – A web address or URL

2 - A “verb”

3 - User Agent

HTTP Request: Verb

GET – Retrieves data

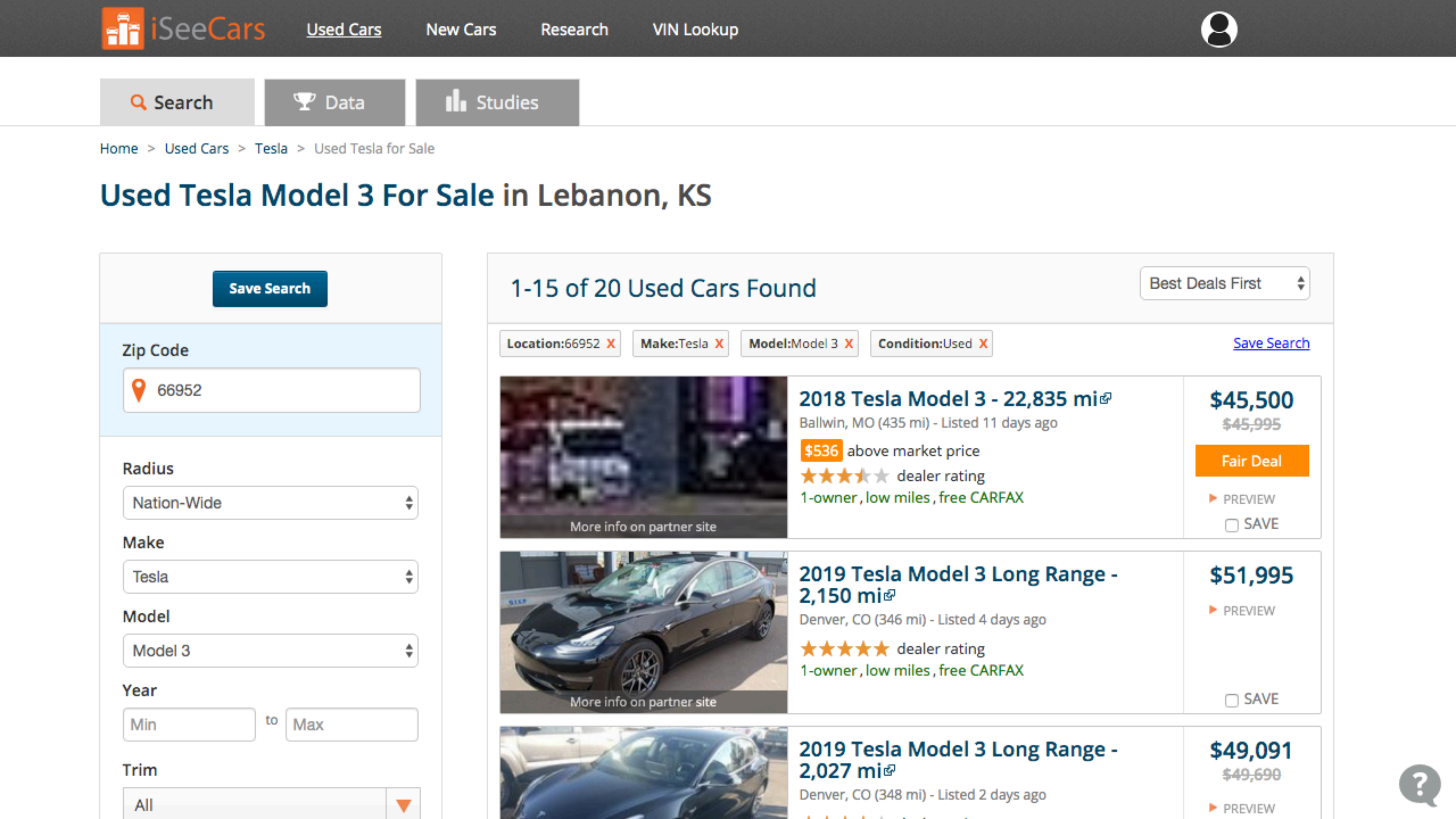
POST – Sends data to the server

HTTP Request: User Agent

Identifies the browser or web scraper

Mozilla/5.0 (Macintosh; Intel Mac OS X 10_13_6) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/78.0.3904.87 Safari/537.36

URL Hacking



Home > Used Cars > Tesla > Used Tesla for Sale

Used Tesla Model 3 For Sale in Lebanon, KS

Zip Code: 66952

Radius: Nation-Wide

Make: Tesla

Model: Model 3

Year: Min to Max

Trim: All

1-15 of 20 Used Cars Found

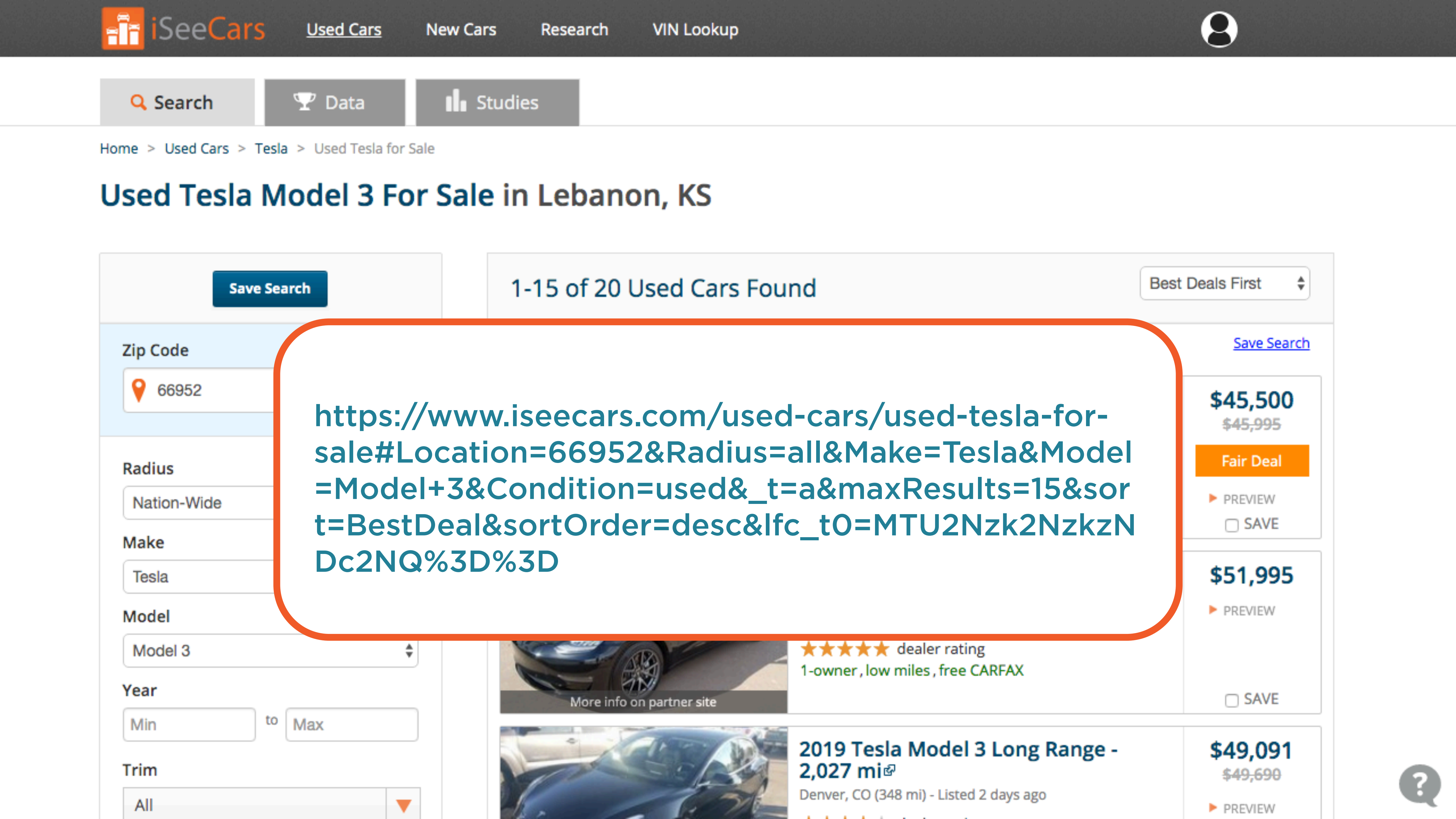
Best Deals First

Location:66952 X | Make:Tesla X | Model:Model 3 X | Condition:Used X

2018 Tesla Model 3 - 22,835 mi
Ballwin, MO (435 mi) - Listed 11 days ago
$536 above market price
★★★⯪☆ dealer rating
1-owner , low miles , free CARFAX
More info on partner site
$45,500
~~$45,995~~
Fair Deal
▶ PREVIEW
□ SAVE

2019 Tesla Model 3 Long Range - 2,150 mi
Denver, CO (346 mi) - Listed 4 days ago
★★★★★ dealer rating
1-owner , low miles , free CARFAX
More info on partner site
$51,995
▶ PREVIEW
□ SAVE

2019 Tesla Model 3 Long Range - 2,027 mi
Denver, CO (348 mi) - Listed 2 days ago
$49,091
~~$49,690~~
▶ PREVIEW



Home > Used Cars > Tesla > Used Tesla for Sale

Used Tesla Model 3 For Sale in Lebanon, KS

Zip Code: 66952

Radius: Nation-Wide

Make: Tesla

Model: Model 3

Year: Min to Max

Trim: All

1-15 of 20 Used Cars Found

https://www.iseecars.com/used-cars/used-tesla-for-sale#Location=66952&Radius=all&Make=Tesla&Model=Model+3&Condition=used&_t=a&maxResults=15&sort=BestDeal&sortOrder=desc&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D

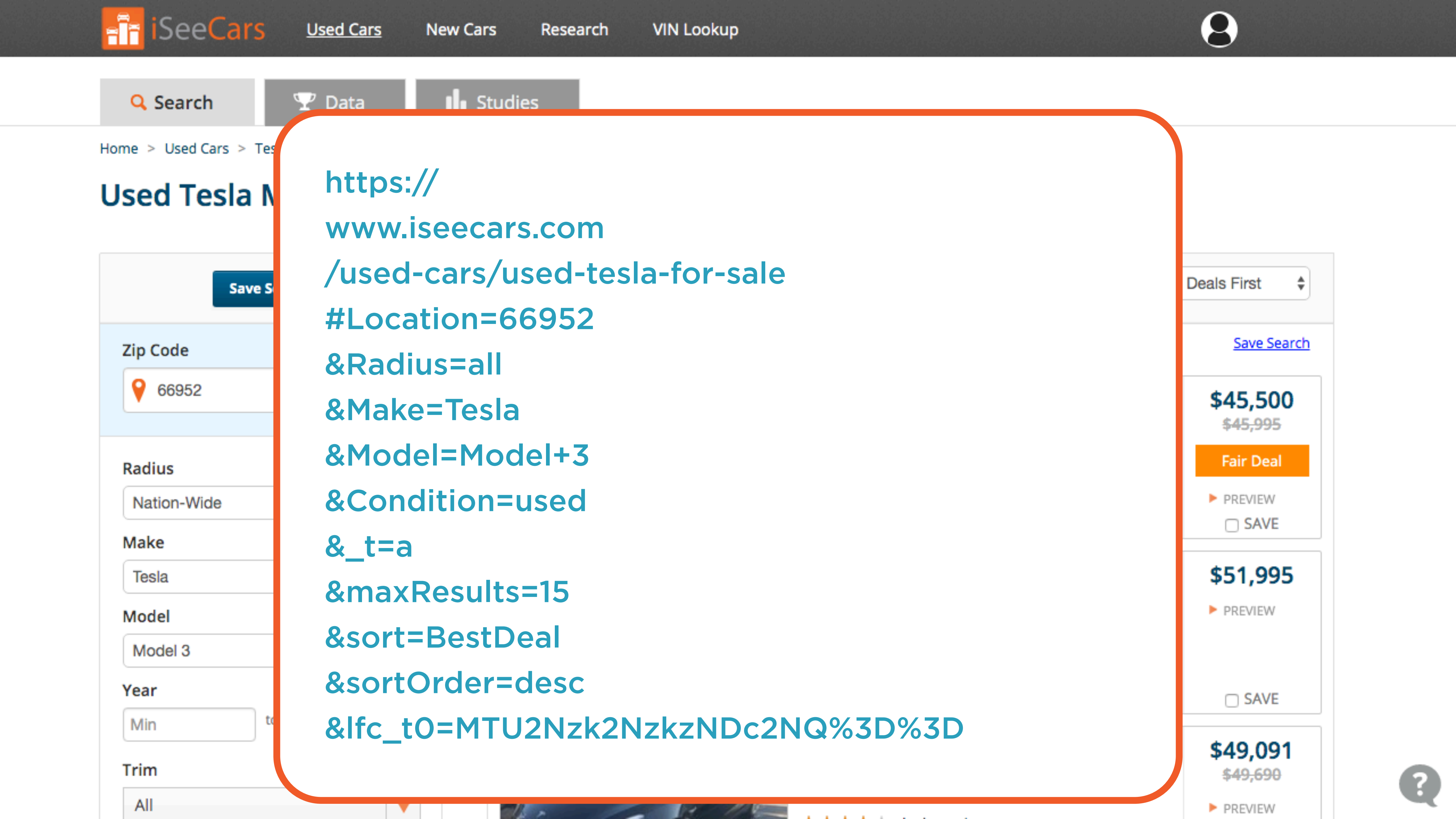
$45,500
~~$45,995~~
Fair Deal

$51,995

★★★★★ dealer rating
1-owner, low miles, free CARFAX

2019 Tesla Model 3 Long Range - 2,027 mi
Denver, CO (348 mi) - Listed 2 days ago

$49,091
~~$49,690~~



Used Tesla M

Zip Code
66952

Radius
Nation-Wide

Make
Tesla

Model
Model 3

Year
Min to

Trim
All

$45,500
~~$45,995~~
Fair Deal

$51,995

$49,091
~~$49,690~~

https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D

Scheme

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

iseescars.com URL

Host

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Port

iseescars.com URL

```
https://
www.iseecars.com:443
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Path

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String (?)
or
URL Fragment (#)

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

Query String

iseescars.com URL

```
https://
www.iseecars.com
/used-cars/used-tesla-for-sale
#Location=66952
&Radius=all
&Make=Tesla
&Model=Model+3
&Condition=used
&_t=a
&maxResults=15
&sort=BestDeal
&sortOrder=desc
&lfc_t0=MTU2Nzk2NzkzNDc2NQ%3D%3D
```

```
host = ‘www.iseecars.com'
path = ‘/used-cars/used-tesla-for-sale’
location = ‘66952’
query_string = f’#Location={location}&Radius=all&Make=Tesla&Model=Model+3’

start_url = f’http://{host}{path}{query_string}’
```

Python URL Strings

```
import requests
start_url = ‘https://www.iseecars.com/used-cars/used-tesla-for-sale’

downloaded_page = requests.get(start_url)

print(downloaded_page.text)
```

Python Requests

Summary

Human browsing versus web scraping

HTTP protocol

URL hacking