Lecture 18 | Web Scraping

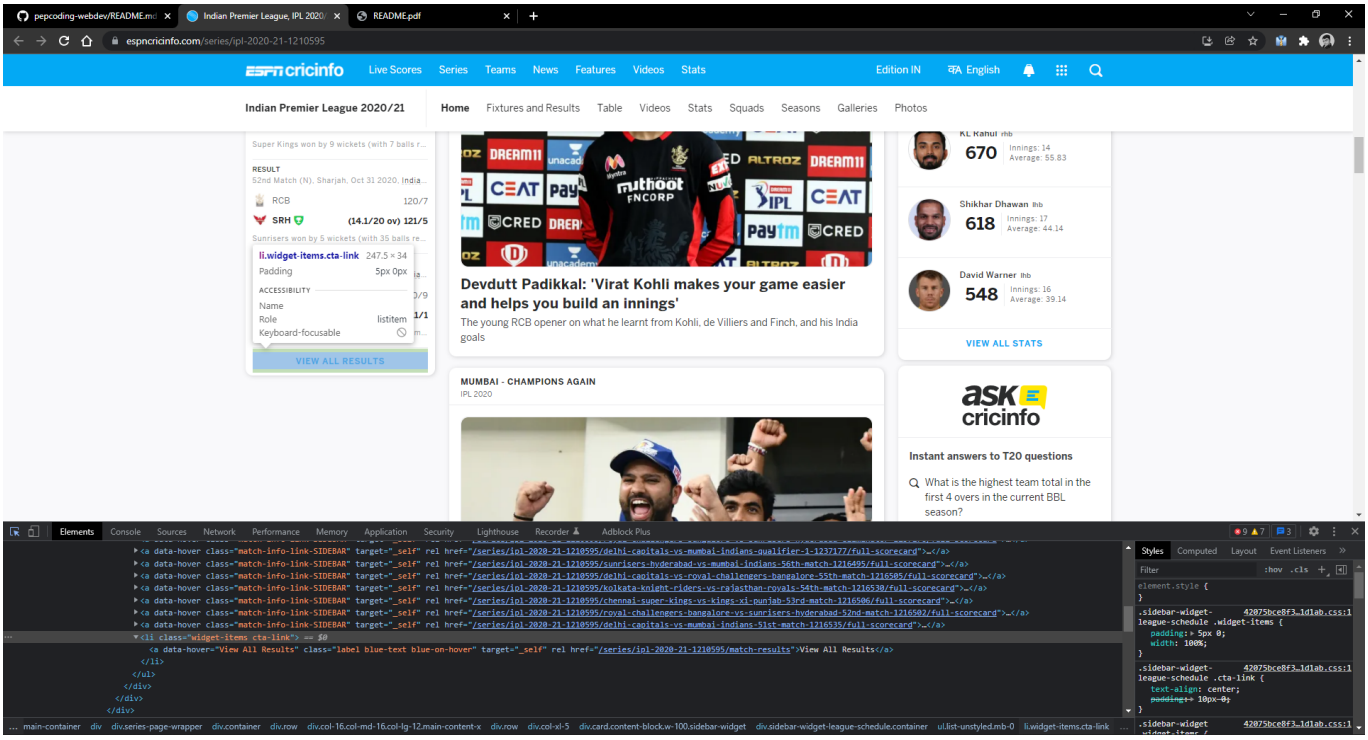
! ESPN Web Scraping Project

> ESPN Cricinfo | https://www.espncricinfo.com/series/ipl-2020-21-1210595

Setup the folder structure

- Make a folder named `espn-scraping`
- Inside the `espn-scraping` folder, `npm init -y` : it will create a `package.json` file
- `npm i cheerio request` in the folder

> The actual **anchor** tag to the `scoreboard`



> Code to *scrape the data* from the website (view all results)

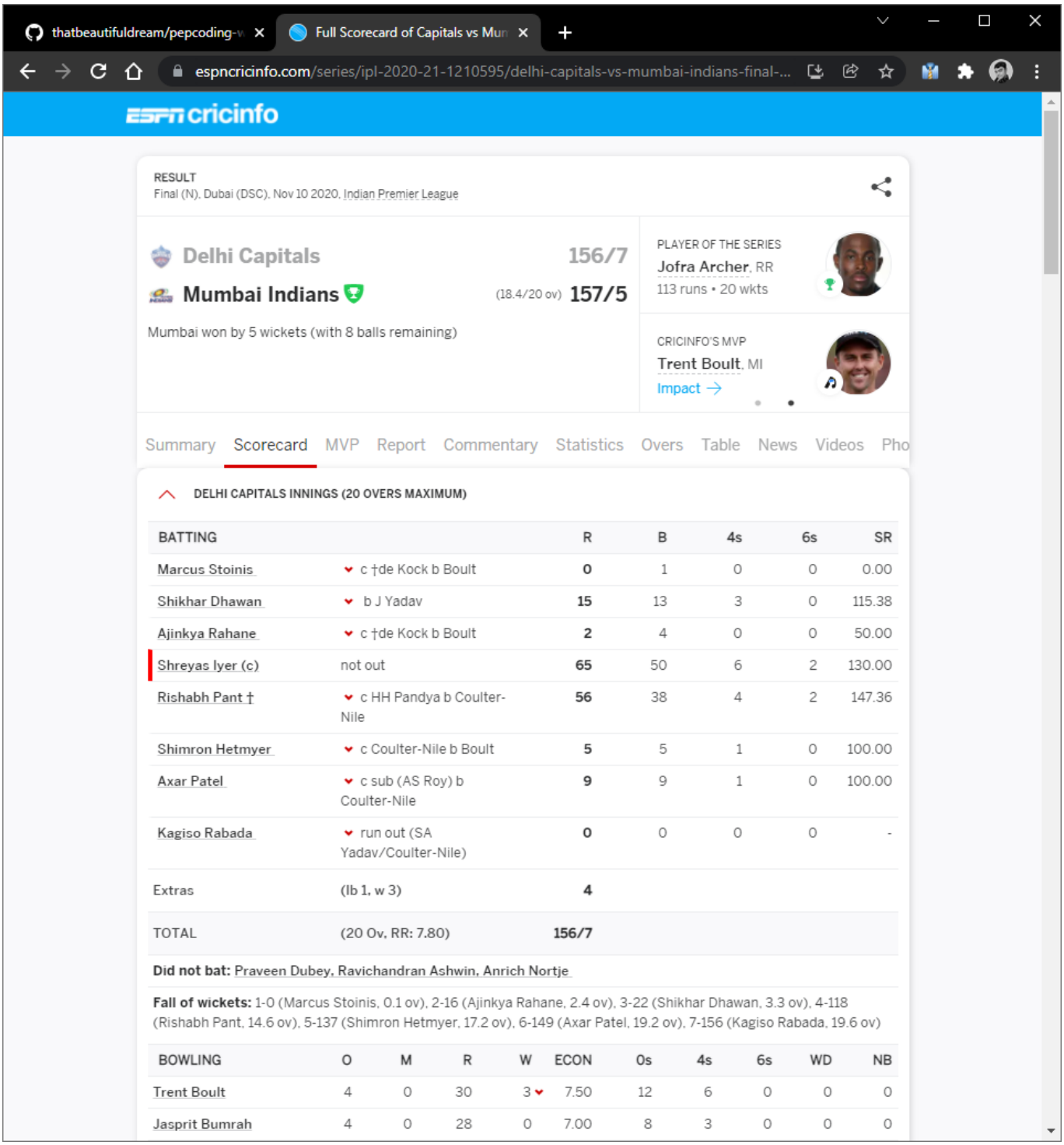
```
const url = "https://www.espncricinfo.com/series/ipl-2020-21-1210595";
const cheerio = require("cheerio");
const request = require("request");

request(url, cb);

// cb is a callback function for request
function cb(err, response, html) {
  if (err) {
    console.error(err);
  } else {
    extractLink(html);
  }
```

```
}

function extractLink(html) {
  const $ = cheerio.load(html);
  let anchorElement = $("a[data-hover = 'View All Results']");
  let link = anchorElement.attr("href");
  let fullLink = "https://www.espncricinfo.com" + link;
  console.log(fullLink);
}
```

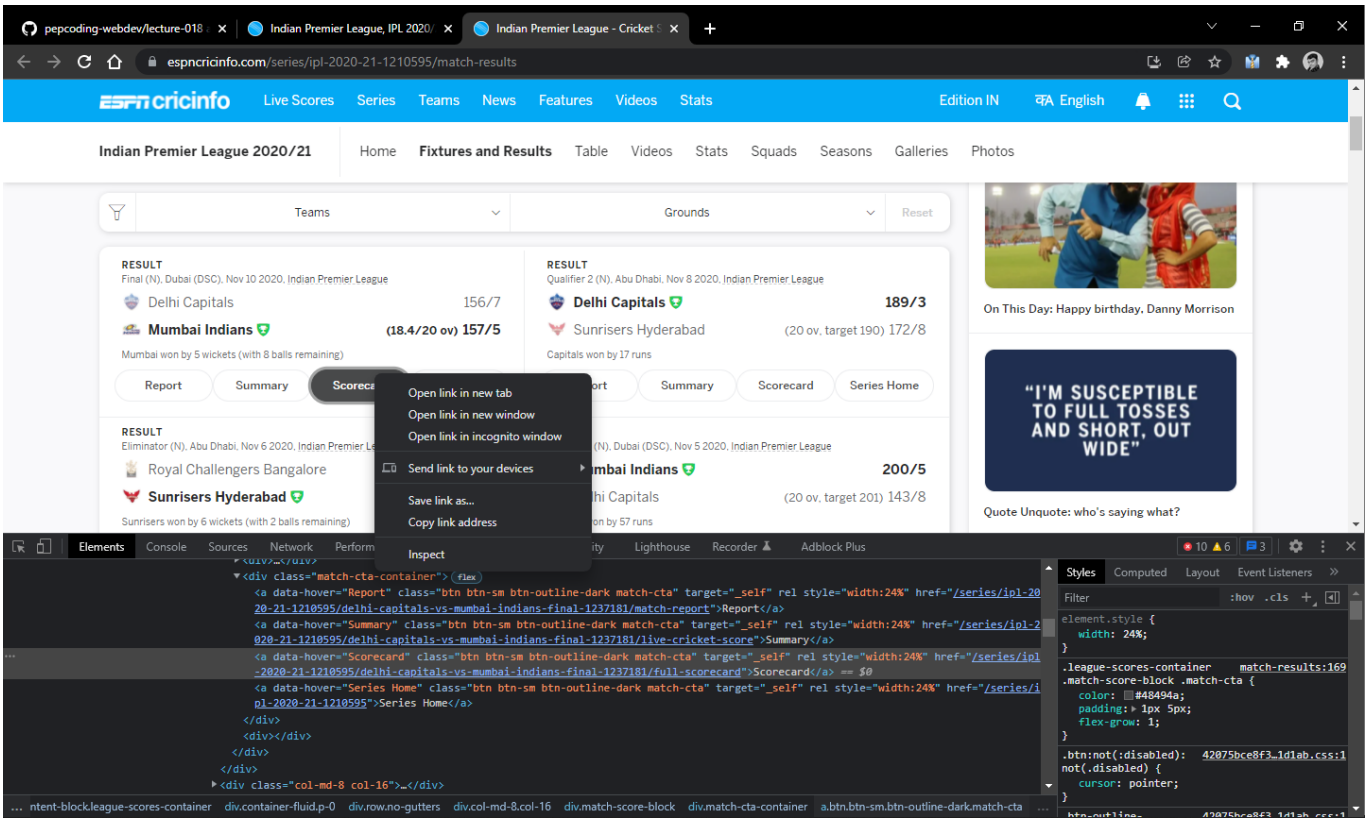
> Output :

```
$ node main.js
https://www.espncricinfo.com/series/ipl-2020-21-1210595/match-results
```

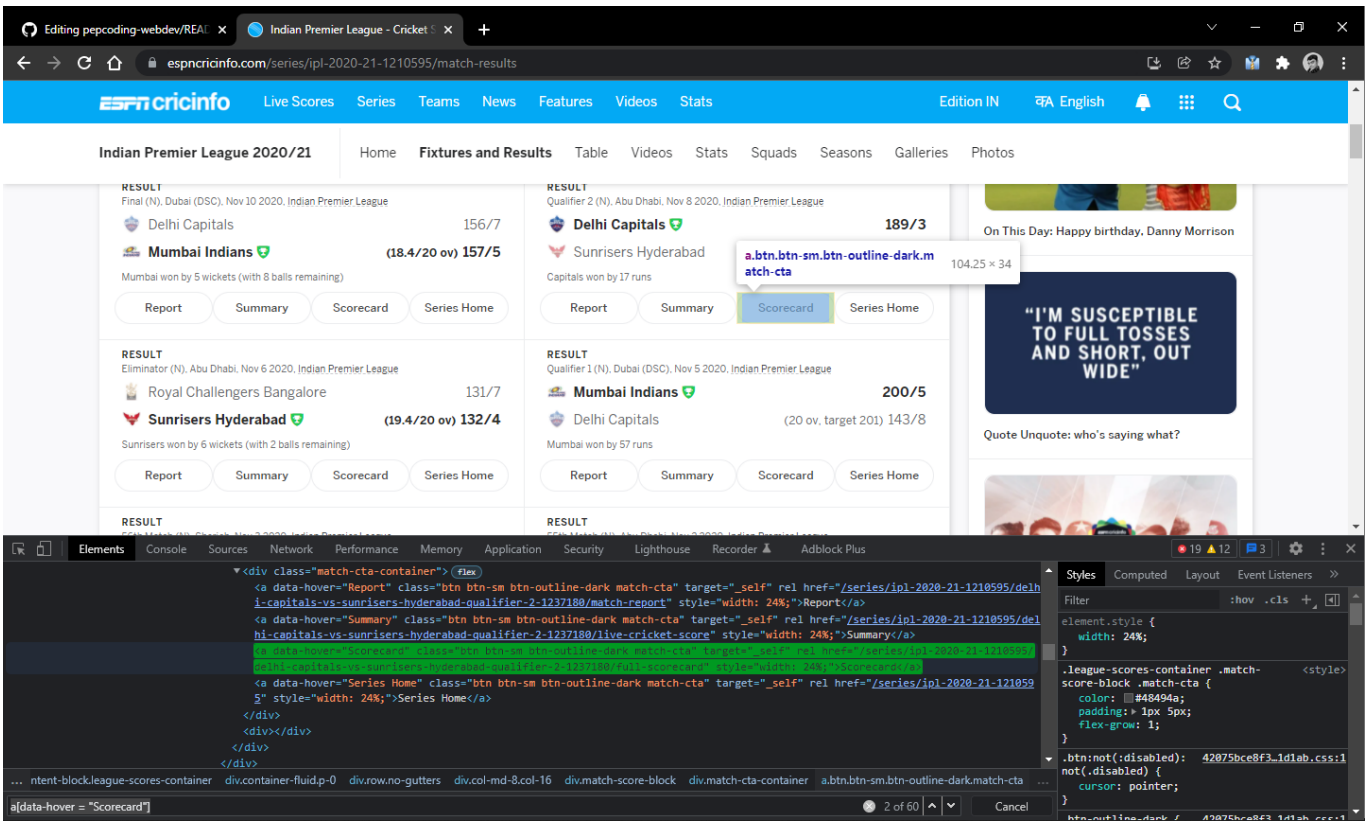
> The extracted `view all results` link



> Scorecard anchor



Scorecard `a[data-hover = "Scorecard"]` array



Extracting all the `scorecard` links

```
const url = "https://www.espncricinfo.com/series/ipl-2020-21-1210595";
const cheerio = require("cheerio");
const request = require("request");
```

```
request(url, cb);

// cb is a callback function for request
function cb(err, response, html) {
  if (err) {
    console.error(err);
  } else {
    extractLink(html);
  }
}

function extractLink(html) {
  const $ = cheerio.load(html);
  let anchorElement = $("a[data-hover = 'View All Results']");
  let link = anchorElement.attr("href");
  let fullLink = "https://www.espncricinfo.com" + link; // view all results
  console.log(fullLink);

  getAllMatchLink(fullLink);
}

function getAllMatchLink(uri) {
  request(uri, function (err, response, html) {
    if (err) {
      console.error(err);
    } else {
      extractAllLink(html); // all scorecard link
    }
  });
}

function extractAllLink(html) {
  const $ = cheerio.load(html);
  let scoreCardArr = $('a[data-hover = "Scorecard"]');
  for (let i = 0; i < scoreCardArr.length; i++) {
    let link = scoreCardArr[i].attribs.href;
    let fullLink = "https://www.espncricinfo.com" + link;
    console.log(fullLink);
  }
}
```

Output :

```
$ node main.js
https://www.espncricinfo.com/series/ipl-2020-21-1210595/match-results
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-mumbai-
indians-final-1237181/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-
sunrisers-hyderabad-qualifier-2-1237180/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-sunrisers-hyderabad-eliminator-1237178/full-scorec
ard
```

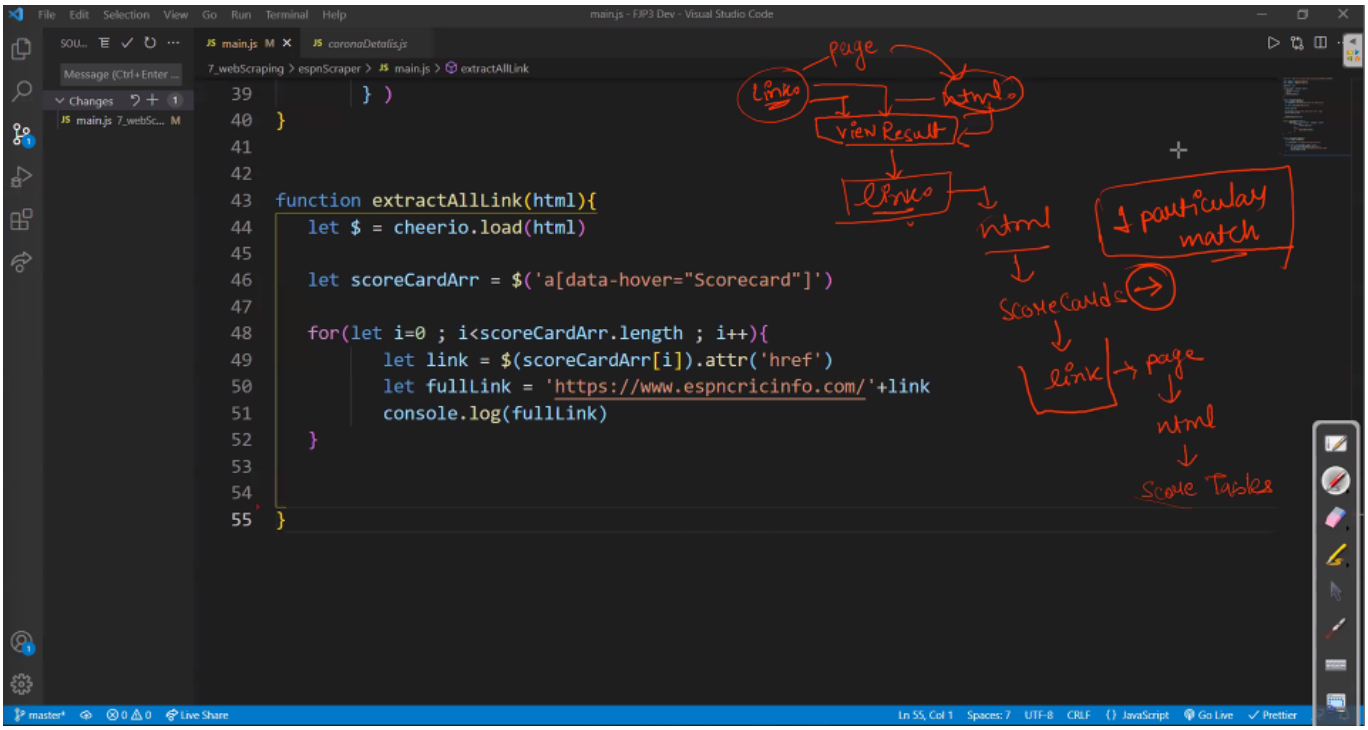
```
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-mumbai-
indians-qualifier-1-1237177/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
mumbai-indians-56th-match-1216495/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-royal-
challengers-bangalore-55th-match-1216505/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
rajasthan-royals-54th-match-1216530/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
kings-xi-punjab-53rd-match-1216506/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-sunrisers-hyderabad-52nd-match-1216502/full-scorec
ard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-mumbai-
indians-51st-match-1216535/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-
rajasthan-royals-50th-match-1216537/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
kolkata-knight-riders-49th-match-1216536/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-royal-
challengers-bangalore-48th-match-1216499/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
delhi-capitals-47th-match-1216524/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
kings-xi-punjab-46th-match-1216520/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-
mumbai-indians-45th-match-1216541/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-chennai-super-kings-44th-match-1216544/full-scorec
ard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-
sunrisers-hyderabad-43rd-match-1216498/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
delhi-capitals-42nd-match-1216497/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
mumbai-indians-41st-match-1216521/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-
sunrisers-hyderabad-40th-match-1216518/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
royal-challengers-bangalore-39th-match-1216494/full-scor
ecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-delhi-
capitals-38th-match-1216546/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
rajasthan-royals-37th-match-1216533/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-kings-
xi-punjab-36th-match-1216517/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
kolkata-knight-riders-35th-match-1216512/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-chennai-
super-kings-34th-match-1216509/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-royal-
challengers-bangalore-33rd-match-1216522/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-kolkata-
```

```
knight-riders-32nd-match-1216526/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-kings-xi-punjab-31st-match-1216531/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-
rajasthan-royals-30th-match-1216543/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
chennai-super-kings-29th-match-1216528/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-kolkata-knight-riders-28th-match-1216540/full-scor
ecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-delhi-
capitals-27th-match-1216529/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
rajasthan-royals-26th-match-1216507/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
royal-challengers-bangalore-25th-match-1216525/full-scorec
ard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-
kolkata-knight-riders-24th-match-1216523/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-delhi-
capitals-23rd-match-1216500/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
kings-xi-punjab-22nd-match-1216542/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
chennai-super-kings-21st-match-1216501/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-
rajasthan-royals-20th-match-1216511/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-delhi-capitals-19th-match-1216519/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-
chennai-super-kings-18th-match-1216513/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-
sunrisers-hyderabad-17th-match-1216538/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-kolkata-
knight-riders-16th-match-1216515/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-rajasthan-royals-15th-match-1216514/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
sunrisers-hyderabad-14th-match-1216516/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-mumbai-
indians-13th-match-1216503/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-
kolkata-knight-riders-12th-match-1216504/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-
sunrisers-hyderabad-11th-match-1216532/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/royal-challengers-
bangalore-vs-mumbai-indians-10th-match-1216547/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-kings-
xi-punjab-9th-match-1216527/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
sunrisers-hyderabad-8th-match-1216545/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/chennai-super-kings-vs-
delhi-capitals-7th-match-1216539/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kings-xi-punjab-vs-royal-
```

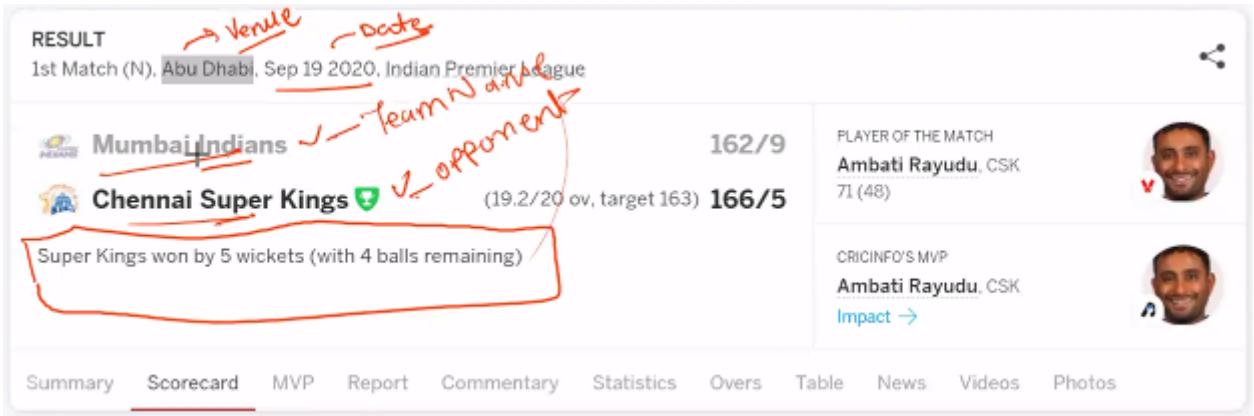
```
challengers-bangalore-6th-match-1216510/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/kolkata-knight-riders-vs-
mumbai-indians-5th-match-1216508/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/rajasthan-royals-vs-
chennai-super-kings-4th-match-1216496/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/sunrisers-hyderabad-vs-
royal-challengers-bangalore-3rd-match-1216534/full-scoreca
rd
https://www.espncricinfo.com/series/ipl-2020-21-1210595/delhi-capitals-vs-kings-
xi-punjab-2nd-match-1216493/full-scorecard
https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-chennai-
super-kings-1st-match-1216492/full-scorecard
```

The `scorecard` page is a `table` with `tr` and `td`

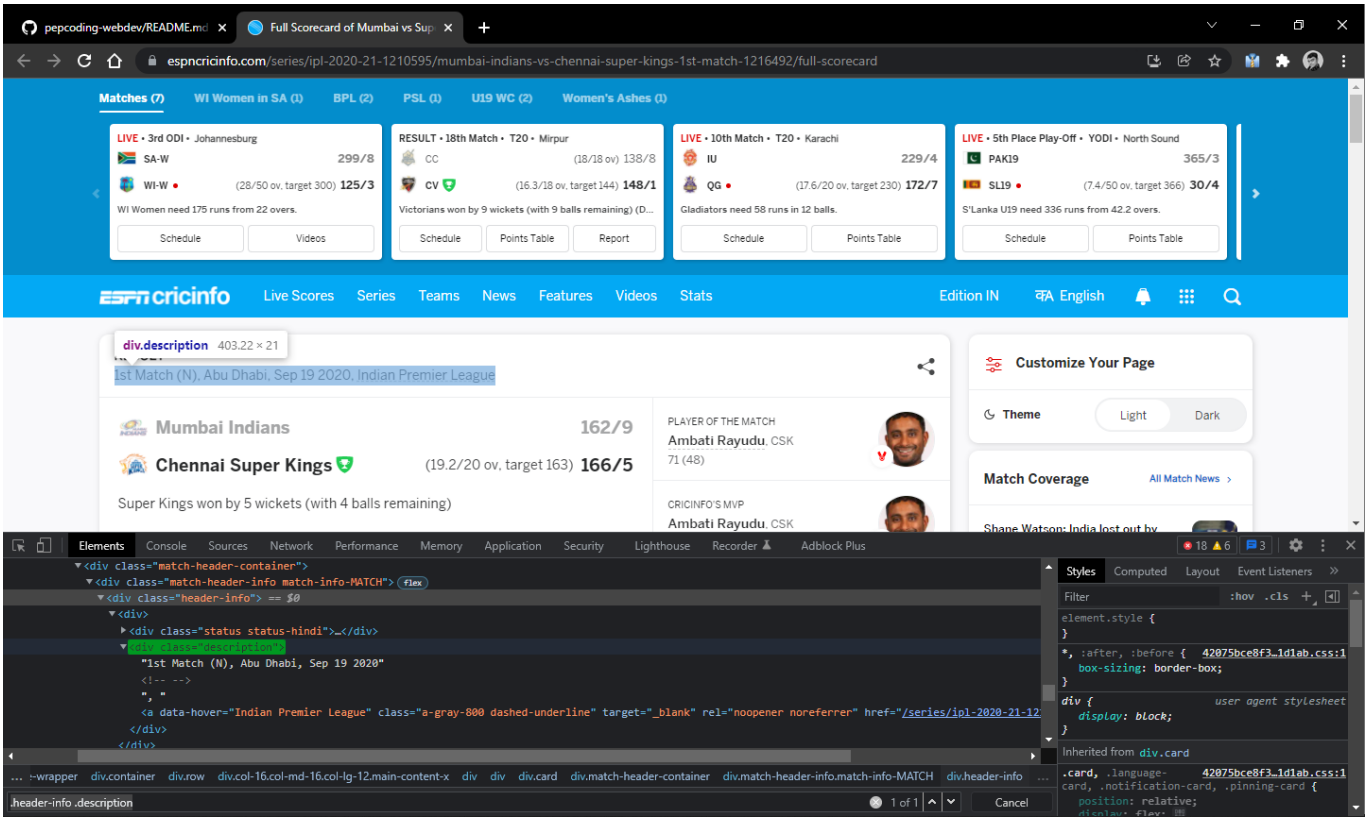
> The idea :



> Task : Scrape following `data` from the `scorecard` page



> idea for the task was to use `split()` on the `string` and then use a for loop to iterate through the array




```
// idea for the task was to use split() on the string and then use a for loop to
iterate through the array

const url =
  "https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-
chennai-super-kings-1st-match-1216492/full-scorecard";

const cheerio = require("cheerio");
const request = require("request");

request(url, cb);

function cb(error, response, html) {
  if (error) {
    console.error(error);
  } else {
    extractMatchDetails(html);
  }
}

function extractMatchDetails(html) {
  const $ = cheerio.load(html);
  let descString = $(".header-info .description").text();
  let descStringArr = descString.split(",");
  console.log(descStringArr);
}
```

> Output :

```
$ node scorecard.js
[
  '1st Match (N)',
  ' Abu Dhabi',
  ' Sep 19 2020',
  ' Indian Premier League'
]
```

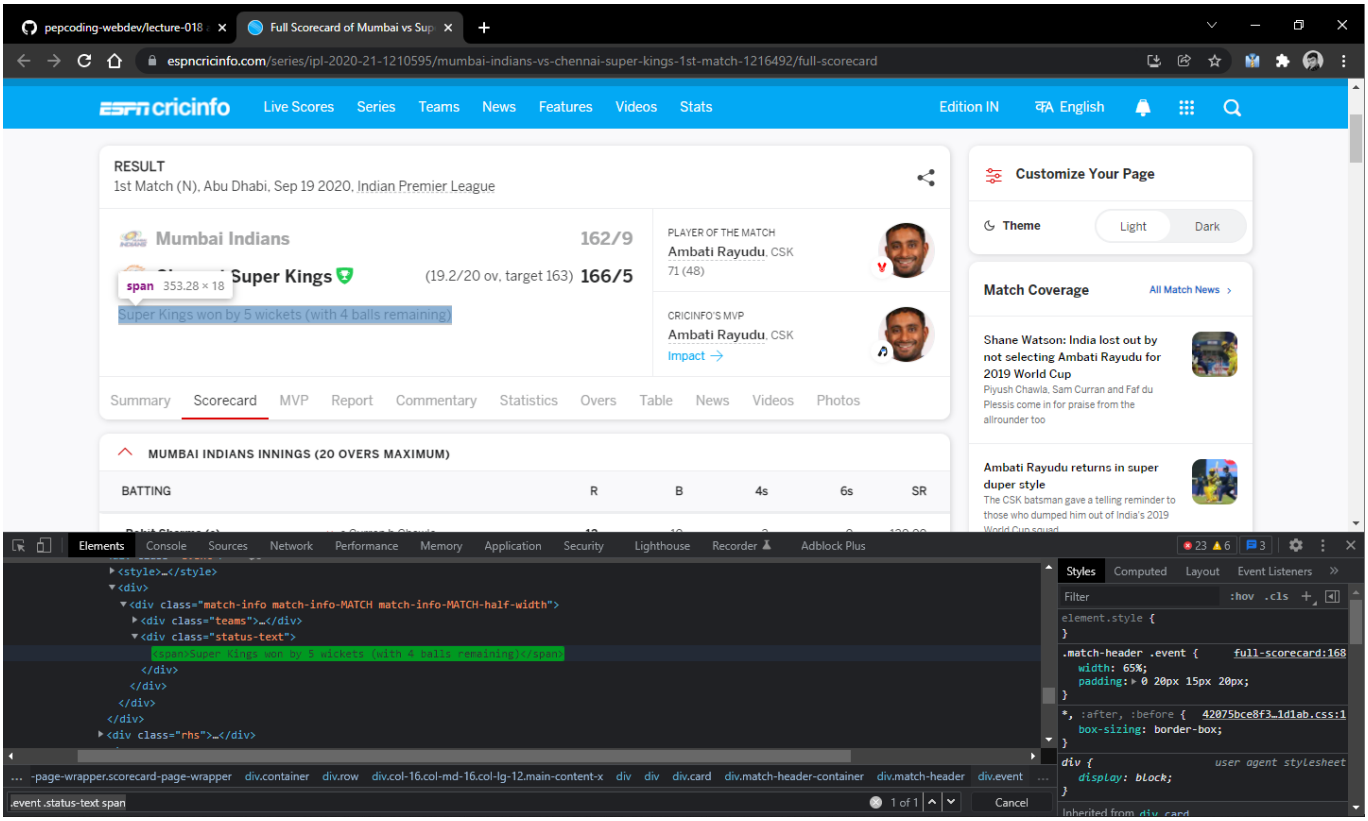
> Improvement : Added `venue`, `date` and `series name` to the output

```
// idea for the task was to use split() on the string and then use a for loop to
iterate through the array

const url =
  "https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-
chennai-super-kings-1st-match-1216492/full-scorecard";

const cheerio = require("cheerio");
const request = require("request");

request(url, cb);

function cb(error, response, html) {
  if (error) {
    console.error(error);
  } else {
    extractMatchDetails(html);
  }
}

function extractMatchDetails(html) {
  const $ = cheerio.load(html);
  let descString = $(".header-info .description").text();
  let descStringArr = descString.split(",");
  // console.log(descStringArr);
  let venue = descStringArr[1].trim(); // using trim() to remove the white space
is a good practice
  let date = descStringArr[2].trim();
  let matchType = descStringArr[3].trim();
  console.log(venue);
  console.log(date);
  console.log(matchType);
}
```

```
$ node scorecard.js
Abu Dhabi
```

```
Sep 19 2020
Indian Premier League
```

Get result text



- Note : To select the desired class use parent class and then child class (desendent)

```
// idea for the task was to use split() on the string and then use a for loop to
iterate through the array

const url =
  "https://www.espncricinfo.com/series/ipl-2020-21-1210595/mumbai-indians-vs-
chennai-super-kings-1st-match-1216492/full-scorecard";

const cheerio = require("cheerio");
const request = require("request");

request(url, cb);

function cb(error, response, html) {
  if (error) {
    console.error(error);
  } else {
    extractMatchDetails(html);
  }
}

function extractMatchDetails(html) {
  const $ = cheerio.load(html);
```

```
  let descString = $(".header-info .description").text();
  let descStringArr = descString.split(",");
  // console.log(descStringArr);
  let venue = descStringArr[1].trim(); // using trim() to remove the white space
is a good practice
  let date = descStringArr[2].trim();
  let matchType = descStringArr[3].trim();
  let results = $(".event .status-text span").text();
  console.log(`Match Details: ${venue}, ${date}, ${matchType}`);
  console.log(`Match Results: ${results}`);
}
```

> Output :

```
$ node scorecard.js
Match Details: Abu Dhabi, Sep 19 2020, Indian Premier League
Match Results: Super Kings won by 5 wickets (with 4 balls remaining)
```

> Next Class : Retrive Team Name, get batsman details and repeat the process for all the matches, and explore the excel module to create a spreadsheet with the data