In [1]:

```
import os

SEASONS = list(range(2020,2024))

DATA_DIR = "data"
STANDINGS_DIR = os.path.join(DATA_DIR, "standings")
SCORES_DIR = os.path.join(DATA_DIR, "scores")
```

In [2]:

```
from bs4 import BeautifulSoup
from playwright.async_api import async_playwright, TimeoutError as Playwrigh
import time
```

In [3]:

```
SEASONS
```

Out[3]:

```
[2020, 2021, 2022, 2023]
```

In [4]:

```
async def get_html(url, selector, sleep=2, retries=3):
    html = None
    for i in range(1, retries+1):
        time.sleep(sleep * i)
        try:
            async with async_playwright() as p:
                browser = await p.firefox.launch()
                page = await browser.new_page()
                await page.goto(url)
                print(await page.title())
                html = await page.inner_html(selector)
        except PlaywrightTimeout:
            print(f"Timeout error on {url}")
            continue
        else:
            break
    return html
```

In [5]:

```
async def scrape_season(season):
    url = f"https://www.basketball-reference.com/leagues/NBA_{season}_games.
    html = await get_html(url, "#content .filter")

    soup = BeautifulSoup(html)
    links = soup.find_all("a")
    standings_pages = [f"https://www.basketball-reference.com{l['href']}" fo

    for url in standings_pages:
        save_path = os.path.join(STANDINGS_DIR, url.split("/")[-1])
        if os.path.exists(save_path):
            continue

        html = await get_html(url, "#all_schedule")
        with open(save_path, "w+") as f:
            f.write(html)
```

In [10]:

```
for season in SEASONS:
```

```python
    await scrape_season(season)
```

```
2019-20 NBA Schedule | Basketball-Reference.com
2020-21 NBA Schedule | Basketball-Reference.com
2021-22 NBA Schedule | Basketball-Reference.com
2022-23 NBA Schedule | Basketball-Reference.com
```

In [6]:

```python
standings_files = os.listdir(STANDINGS_DIR)
```

In [7]:

```python
async def scrape_game(standings_file):
    with open(standings_file, 'r') as f:
        html = f.read()

    soup = BeautifulSoup(html)
    links = soup.find_all("a")
    hrefs = [l.get('href') for l in links]
    box_scores = [f"https://www.basketball-reference.com{l}" for l in hrefs

    for url in box_scores:
        save_path = os.path.join(SCORES_DIR, url.split("/")[-1])
        if os.path.exists(save_path):
            continue

        html = await get_html(url, "#content")
        if not html:
            continue
        with open(save_path, "w+") as f:
            f.write(html)
```

In [13]:

```python
import pandas as pd

for season in SEASONS:
    files = [s for s in standings_files if str(season) in s]

    for f in files:
        filepath = os.path.join(STANDINGS_DIR, f)

        await scrape_game(filepath)
```

In [ ]:

In [ ]: