RED BUS DATA SCRAPING - CODE EXPLANATION

Web Scraping for One Transportation is explained in detail here. Similarly it is done for all 10 Transports.

1.IMPORTING LIBRARIES:

```
from selenium import webdriver
from selenium.webdriver.common.by import By
from selenium.webdriver.common.keys import Keys
from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver.support.ui import WebDriverWait
from selenium.webdriver.support import expected_conditions as EC
import time
import pandas as pd
```

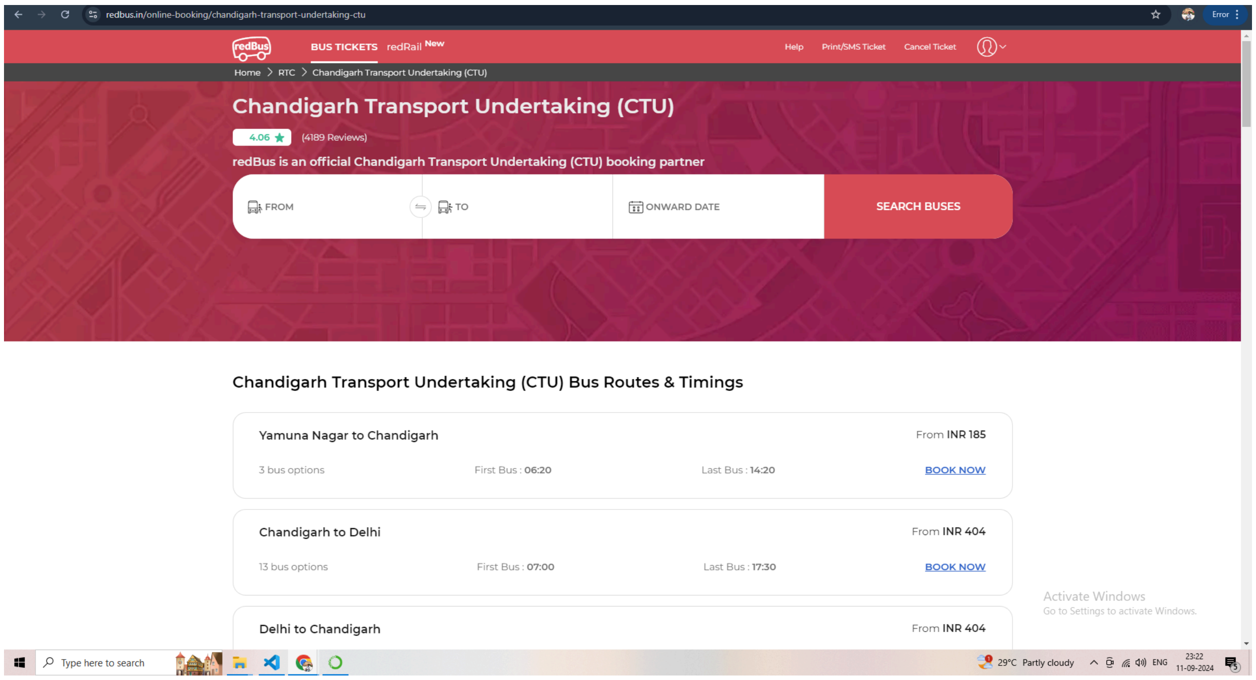
- **Selenium.webdriver:** controls the browser through chrome web driver.
- **By:** Locates elements on a web page via XPATH, ID etc,.
- **Keys**: used for making keyboard simulations
- **ActionChains:** Used for advanced and continuous set of automated actions
- **WebDriverWait :** Waits for an event/ action to take place
- **Expected_conditions:** gives conditions for when to perform an action
- **Time:** Delays an action for given period of time
- **Pandas:** used to create data frames and to save as csv files.

2.Opening URL:

```
# Initializing the Chrome driver
driver = webdriver.Chrome()

# Opening the desired URL
driver.get('https://www.redbus.in/online-booking/chandigarh-transport-unde
rtaking-ctu')
```

```
# Maximizing the browser window
driver.maximize_window()
```

webdriver.Chrome() : Opens Chrome web browser, driver.get(url) connects to the given url and ***driver.maximize_window()*** maximizes the window to full screen.


redbus.in/online-booking/chandigarh-transport-undertaking-ctu
BUS TICKETS
redRail New
Help
Print/SMS Ticket
Cancel Ticket
Home > RTC > Chandigarh Transport Undertaking (CTU)
FROM
TO
ONWARD DATE
SEARCH BUSES
BOOK NOW
Type here to search

Chandigarh Transport Undertaking (CTU)

4.06 ★ (4189 Reviews)

redBus is an official Chandigarh Transport Undertaking (CTU) booking partner

Chandigarh Transport Undertaking (CTU) Bus Routes & Timings

Yamuna Nagar to Chandigarh — From INR 185 — 3 bus options — First Bus : 06:20 — Last Bus : 14:20 — BOOK NOW

Chandigarh to Delhi — From INR 404 — 13 bus options — First Bus : 07:00 — Last Bus : 17:30 — BOOK NOW

Delhi to Chandigarh — From INR 404

2.Function to Scrape Links (hrefs) and Route Names:

```
# Function to scrape hrefs on the current page
def scrape_hrefs():
    bus_routes_links = driver.find_elements(By.XPATH,
"//div[contains(@class, 'route_details')]//a")
    print(f"Found {len(bus_routes_links)} links")
    hrefs = [link.get_attribute('href') for link in bus_routes_links]
    route_names = [link.text for link in bus_routes_links]
    for href, route in zip(hrefs, route_names):
        print(f"Scraped href: {href}, Route: {route}")
    return hrefs, route_names
```

- ***find_elements():*** finds all elements matching the given conditions , in this case the XPATH
- ***get_attribute():*** extracts URL from the <a> tag.
- ***text():***extracts the text inside the xpath
- ***zip():***combines href and route names

A.Scraping Through the First Page:

```
# List to store all hrefs
all_hrefs = []
all_route_names = []

# Scrolling the page a specific number of times (e.g., twice) on the first
page only
for _ in range(3):
    body = driver.find_element(By.TAG_NAME, "body")
    ActionChains(driver).send_keys(Keys.PAGE_DOWN).perform()
    time.sleep(3)  # Wait for the page to load new content

# Scraping hrefs on the first page
hrefs, route_names = scrape_hrefs()
all_hrefs.extend(hrefs)
all_route_names.extend(route_names)
```

Lists are initialized to store the extracted data, using the for loop scrolling is done with sleeping time of 3 seconds to load the page.

Here the ***find_element()*** uses TAG_NAME

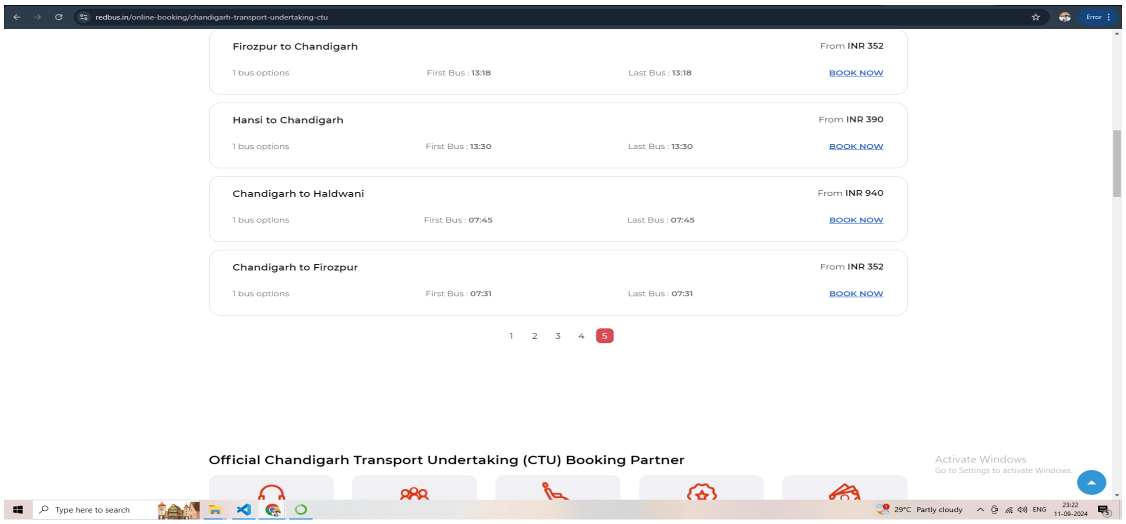
ActionChains are used with ***send_keys()*** to automate scrolling the page by simulating the keyboard “PAGE DOWN” key.

The previously defined function ***scrape_href()*** is used and ***extend()*** is used to append the extracted data to the defined lists.

B.Looping Through the Next Pages:

```python
# Looping through the next 4 pages
for page_num in range(2, 6):
    # Clicking the page number button to go to the next page
    page_button_xpath = f"//div[contains(@class,'DC_117_pageTabs') and
text()='{page_num}']"
    page_button = driver.find_element(By.XPATH, page_button_xpath)
    page_button.click()

    # Waiting for the new page to load
    time.sleep(5)

    # Scraping hrefs on the current page
    hrefs, route_names = scrape_hrefs()
    all_hrefs.extend(hrefs)
    all_route_names.extend(route_names)

# Initializing an empty DataFrame for all routes
CTU = pd.DataFrame()
```

As in the taken example there are 5 pages in total, the loop is initiated for the ***range (2,6)***. After selecting the next page, the data are scraped, appended to the list and looped through all pages.


redbus.in/online-booking/chandigarh-transport-undertaking-ctu
Error
1 2 3 4 5
Activate Windows
Go to Settings to activate Windows.
Type here to search
29°C Partly cloudy
ENG
23:22
11-09-2024

Firozpur to Chandigarh			From INR 352
1 bus options	First Bus : 13:18	Last Bus : 13:18	BOOK NOW
Hansi to Chandigarh			From INR 390
1 bus options	First Bus : 13:30	Last Bus : 13:30	BOOK NOW
Chandigarh to Haldwani			From INR 940
1 bus options	First Bus : 07:45	Last Bus : 07:45	BOOK NOW
Chandigarh to Firozpur			From INR 352
1 bus options	First Bus : 07:31	Last Bus : 07:31	BOOK NOW

Official Chandigarh Transport Undertaking (CTU) Booking Partner

2.Safety Mechanism:

A.EMPTY ROWS:

```
# Function to extract text and handle missing ratings
def extract_text_with_default(elements, default_value='new bus'):
    return [element.text if element.text else default_value for element in
elements]

# Function to pad lists to the maximum length
def pad_list(lst, max_length):
    return lst + [''] * (max_length - len(lst))
```

There are few null values in the appended list , particularly in the rating column. The reason is, those are “New Buses” which are yet to be rated. Hence a function is initiated with a default column for empty rows. Another function ***pad_list()*** is also defined to ensure that all the lists have the same length by adding empty strings to shorter lists. Fortunately our data has no such necessity, still they are added as a good practice.

3.Looping & Extraction Through each Route Link :

A.Looping Through Each HREF:

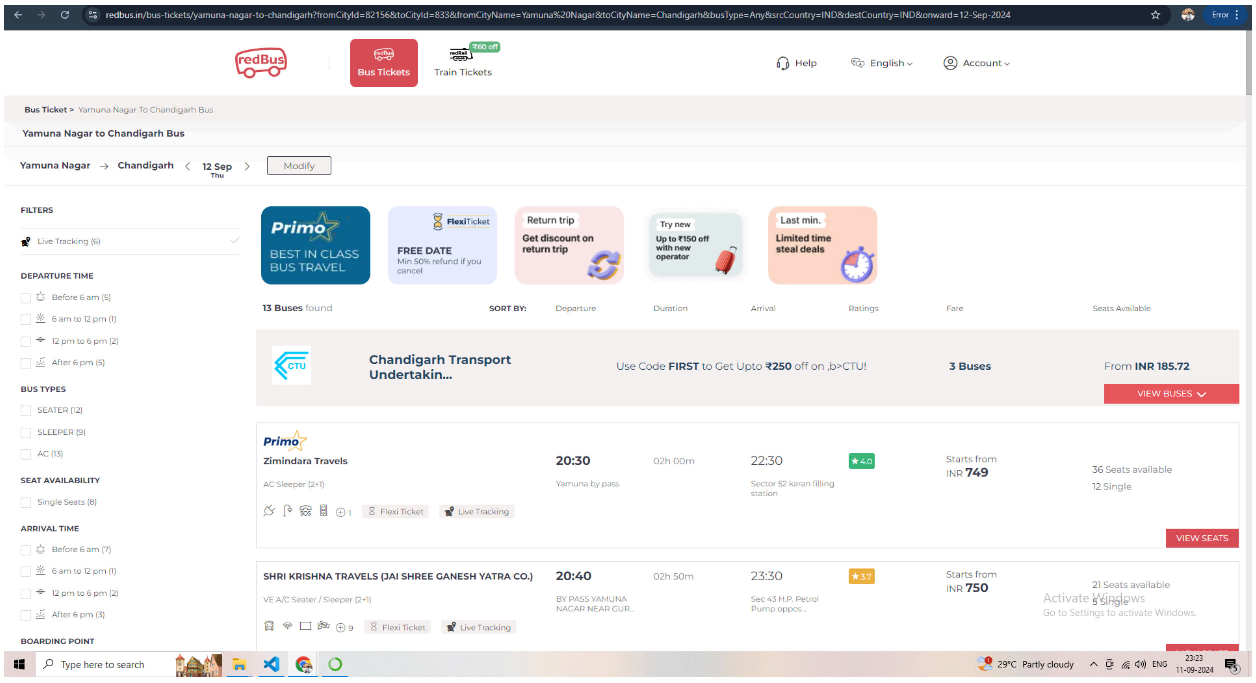
```
# Looping through each href and extract data
for href, route in zip(all_hrefs, all_route_names):
    driver.get(href)

    # Waiting for the page to load
    time.sleep(8)
```

B.Click Buttons if Present:

```
# Clicking all the available buttons (if necessary)
    click_buttons = driver.find_elements(By.XPATH,
"//div[@class='button']")
    for button in click_buttons:
        try:
            WebDriverWait(driver,
10).until(EC.element_to_be_clickable(button)).click()
        except Exception as e:
            print(f"Could not click button: {e}")

    time.sleep(2)
```

In our website there is a necessity to click buttons once every route page is loaded as all the government buses are given inside a dropdown button.

Here Button as a Class is used to find elements and looped through the available buttons.


redbus.in/bus-tickets/yamuna-nagar-to-chandigarh?fromCityId=82156&toCityId=833&fromCityName=Yamuna%20Nagar&toCityName=Chandigarh&busType=Any&srcCountry=IND&destCountry=IND&onward=12-Sep-2024

Bus Ticket > Yamuna Nagar To Chandigarh Bus

Yamuna Nagar to Chandigarh Bus

Yamuna Nagar → Chandigarh 12 Sep Thu

FILTERS
Live Tracking (6)

DEPARTURE TIME
Before 6 am (5)
6 am to 12 pm (1)
12 pm to 6 pm (2)
After 6 pm (5)

BUS TYPES
SEATER (12)
SLEEPER (9)
AC (13)

SEAT AVAILABILITY
Single Seats (8)

ARRIVAL TIME
Before 6 am (7)
6 am to 12 pm (1)
12 pm to 6 pm (2)
After 6 pm (3)

BOARDING POINT

13 Buses found

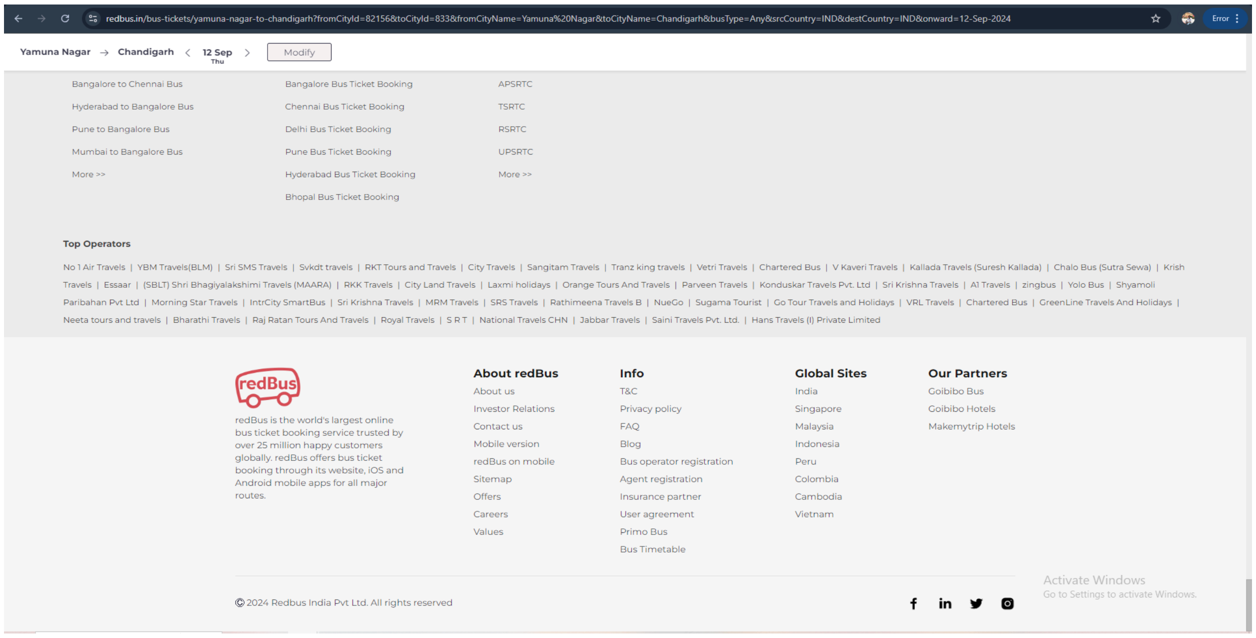
SORT BY:	Departure	Duration	Arrival	Ratings	Fare	Seats Available
Chandigarh Transport Undertakin... — Use Code FIRST to Get Upto ₹250 off on ,b>CTU!					3 Buses	From INR 185.72
Zimindara Travels, AC Sleeper (2+1)	20:30 Yamuna by pass	02h 00m	22:30 Sector 52 karan filling station	★4.0	Starts from INR 749	36 Seats available 12 Single
SHRI KRISHNA TRAVELS (JAI SHREE GANESH YATRA CO.), VE A/C Seater / Sleeper (2+1)	20:40 BY PASS YAMUNA NAGAR NEAR GUR...	02h 50m	23:30 Sec 43 H.P. Petrol Pump oppos...	★3.7	Starts from INR 750	21 Seats available 5 Single

C.Scrolling Through the Page:

```
# Scrolling through the page using ActionChains
    scrolling = True
    max_scroll_attempts = 5  # Maximum number of scroll attempts
    scroll_attempts = 0

    while scrolling and scroll_attempts < max_scroll_attempts:
        old_page_source = driver.page_source

        # Using ActionChains to perform a PAGE_DOWN
        ActionChains(driver).send_keys(Keys.PAGE_DOWN).perform()

        # Waiting for the page to load new content
        time.sleep(3)  # Increased wait time for dynamic content to load

        new_page_source = driver.page_source

        # Checking if the page source has changed
        if new_page_source == old_page_source:
            scroll_attempts += 1
        else:
            scroll_attempts = 0  # Resetting the counter if new content is loaded

        if scroll_attempts >= max_scroll_attempts:
            scrolling = False
```

A False proof logic is used to scroll the page and stop scrolling before the next link is selected for the next loop.

Scroll_attempts is initialized as 0 and then maximum attempts to end the scrolling is initialized as 5. When the scrolling attempt is less than the max scrolling attempt, the pages continue to scroll until new content is loaded. If ***new_page_source*** = ***old_page_source*** then attempt is increased by 1 until it reaches maximum attempts to stop scrolling.


redbus.in/bus-tickets/yamuna-nagar-to-chandigarh?fromCityId=82156&toCityId=833&fromCityName=Yamuna%20Nagar&toCityName=Chandigarh&busType=Any&srcCountry=IND&destCountry=IND&onward=12-Sep-2024

Yamuna Nagar → Chandigarh 12 Sep Thu Modify

Bangalore to Chennai Bus
Hyderabad to Bangalore Bus
Pune to Bangalore Bus
Mumbai to Bangalore Bus
More >>

Bangalore Bus Ticket Booking
Chennai Bus Ticket Booking
Delhi Bus Ticket Booking
Pune Bus Ticket Booking
Hyderabad Bus Ticket Booking
Bhopal Bus Ticket Booking

APSRTC
TSRTC
RSRTC
UPSRTC
More >>

Top Operators

No 1 Air Travels | YBM Travels(BLM) | Sri SMS Travels | Svkdt travels | RKT Tours and Travels | City Travels | Sangitam Travels | Tranz king travels | Vetri Travels | Chartered Bus | V Kaveri Travels | Kallada Travels (Suresh Kallada) | Chalo Bus (Sutra Sewa) | Krish Travels | Essaar | (SBLT) Shri Bhagiyalakshimi Travels (MAARA) | RKK Travels | City Land Travels | Laxmi holidays | Orange Tours And Travels | Parveen Travels | Konduskar Travels Pvt. Ltd | Sri Krishna Travels | A1 Travels | zingbus | Yolo Bus | Shyamoli Paribahan Pvt Ltd | Morning Star Travels | IntrCity SmartBus | Sri Krishna Travels | MRM Travels | SRS Travels | Rathimeena Travels B | NueGo | Sugama Tourist | Go Tour Travels and Holidays | VRL Travels | Chartered Bus | GreenLine Travels And Holidays | Neeta tours and travels | Bharathi Travels | Raj Ratan Tours And Travels | Royal Travels | S R T | National Travels CHN | Jabbar Travels | Saini Travels Pvt. Ltd. | Hans Travels (I) Private Limited

redBus is the world's largest online bus ticket booking service trusted by over 25 million happy customers globally. redBus offers bus ticket booking through its website, iOS and Android mobile apps for all major routes.

About redBus
About us
Investor Relations
Contact us
Mobile version
redBus on mobile
Sitemap
Offers
Careers
Values

Info
T&C
Privacy policy
FAQ
Blog
Bus operator registration
Agent registration
Insurance partner
User agreement
Primo Bus
Bus Timetable

Global Sites
India
Singapore
Malaysia
Indonesia
Peru
Colombia
Cambodia
Vietnam

Our Partners
Goibibo Bus
Goibibo Hotels
Makemytrip Hotels

© 2024 Redbus India Pvt Ltd. All rights reserved

Activate Windows
Go to Settings to activate Windows.

D.Extraction of Data:

```
 # Collecting data
    bus_names = extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='travels lh-24 f-bold d-color']"))
    bus_types = extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='bus-type f-12 m-top-16 l-color evBus']"))
    departure_times =
extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='dp-time f-19 d-color f-bold']"))
    durations = extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='dur l-color lh-24']"))
    reaching_times =
extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='bp-time f-19 d-color disp-Inline']"))
    prices = extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='fare d-block']"))

    rating_elements = driver.find_elements(By.XPATH,
"//div[@class='rating-sec lh-24']")
    ratings = extract_text_with_default(rating_elements)
    if len(ratings) < len(bus_names):
        ratings.extend(['new bus'] * (len(bus_names) - len(ratings)))

    seat_availability =
extract_text_with_default(driver.find_elements(By.XPATH,
"//div[@class='seat-left m-top-16' or @class='seat-left m-top-30']"))
```

The Data bus_name, bus_type, departure_time, duration, reaching_time, prices, ratings, seat_availabilty(while extraction) are extracted using ***find_element()*** and ***extract_text()*** function which is previously defined.

E.Pad The lists:

```
# Finding the maximum length among the lists
    max_length = max(len(bus_names), len(bus_types), len(departure_times),
len(durations), len(reaching_times), len(ratings), len(prices),
len(seat_availability))
```

```
    # Padding all lists to the same length
    bus_names = pad_list(bus_names, max_length)
    bus_types = pad_list(bus_types, max_length)
    departure_times = pad_list(departure_times, max_length)
    durations = pad_list(durations, max_length)
    reaching_times = pad_list(reaching_times, max_length)
    ratings = pad_list(ratings, max_length)
    prices = pad_list(prices, max_length)
    seat_availability = pad_list(seat_availability, max_length)
```

The safety function ***pad-list()*** previously defined is used to check if there is any list with shorter length and filled with empty strings. But fortunately in our data set only Bus Ratings have empty strings due to new buses., which are filled as "New Bus" using a previously defined function. ***Pad_list()*** are done just as a safety mechanism for future usage and as a good practice.

4.Data Frame :

A.creating a DataFrame:

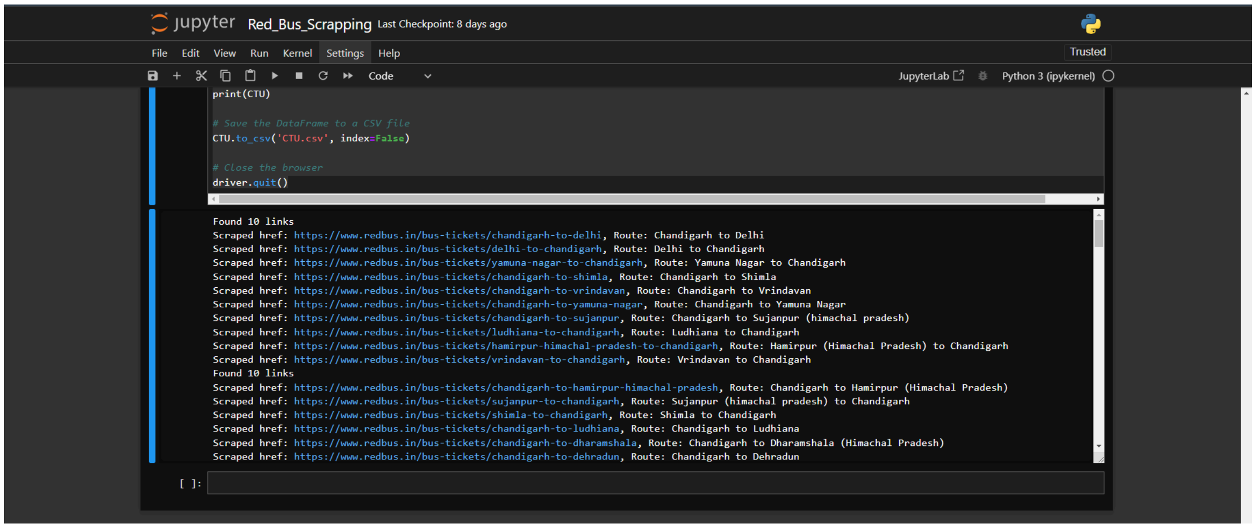
```
# Creating a DataFrame for the current route
    route_df = pd.DataFrame({
        'Route Name': [route] * max_length,
        'Route Link': [href] * max_length,
        'Bus Name': bus_names,
        'Bus Type': bus_types,
        'Departure Time': departure_times,
        'Duration': durations,
        'Reaching Time': reaching_times,
        'Bus Rating': ratings,
        'Price': prices,
        'Seat Availability': seat_availability
    # Concatenating the current route DataFrame with the main DataFrame
    CTU = pd.concat([CTU, route_df], ignore_index=True)
```

A Data Frame is created with all the scraped data from the website and concat together with the route name and route_link df ie,. CTU in this case.

B.Save as a CSV File:

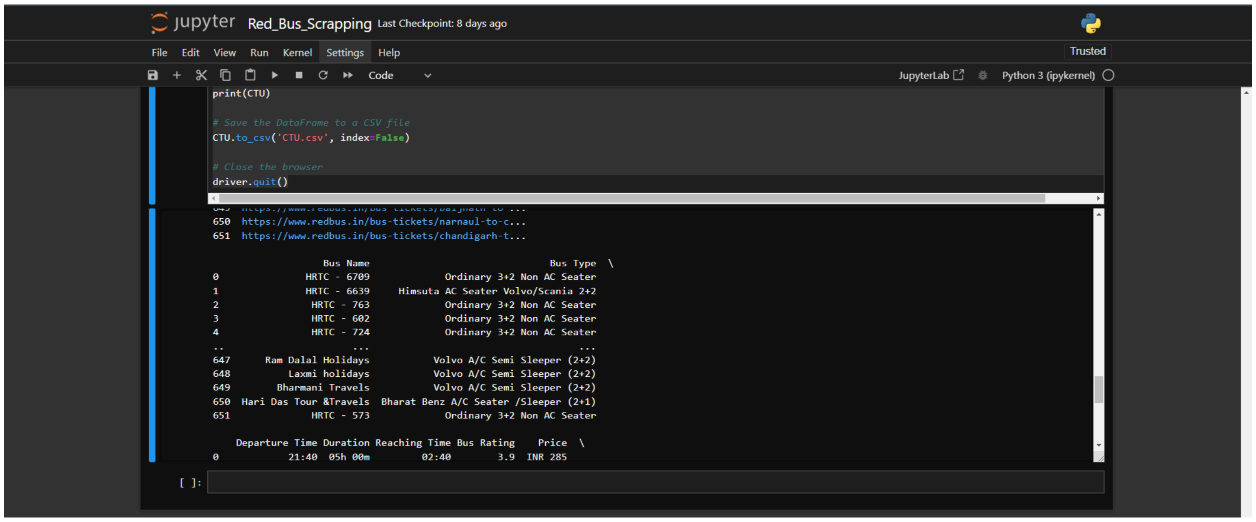
```
# Saving the DataFrame to a CSV file
CTU.to_csv('CTU.csv', index=False)

# Closing the browser
driver.quit()
```

Saved as a CSV file and the web browser is closed automatically after all the extraction and data frame formulation is done.


jupyter
Red_Bus_Scrapping Last Checkpoint: 8 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
[ ]:

```
print(CTU)

# Save the DataFrame to a CSV file
CTU.to_csv('CTU.csv', index=False)

# Close the browser
driver.quit()
```

```
Found 10 links
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-delhi, Route: Chandigarh to Delhi
Scraped href: https://www.redbus.in/bus-tickets/delhi-to-chandigarh, Route: Delhi to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/yamuna-nagar-to-chandigarh, Route: Yamuna Nagar to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-shimla, Route: Chandigarh to Shimla
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-vrindavan, Route: Chandigarh to Vrindavan
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-yamuna-nagar, Route: Chandigarh to Yamuna Nagar
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-sujanpur, Route: Chandigarh to Sujanpur (himachal pradesh)
Scraped href: https://www.redbus.in/bus-tickets/ludhiana-to-chandigarh, Route: Ludhiana to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/hamirpur-himachal-pradesh-to-chandigarh, Route: Hamirpur (Himachal Pradesh) to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/vrindavan-to-chandigarh, Route: Vrindavan to Chandigarh
Found 10 links
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-hamirpur-himachal-pradesh, Route: Chandigarh to Hamirpur (Himachal Pradesh)
Scraped href: https://www.redbus.in/bus-tickets/sujanpur-to-chandigarh, Route: Sujanpur (himachal pradesh) to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/shimla-to-chandigarh, Route: Shimla to Chandigarh
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-ludhiana, Route: Chandigarh to Ludhiana
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-dharamshala, Route: Chandigarh to Dharamshala (Himachal Pradesh)
Scraped href: https://www.redbus.in/bus-tickets/chandigarh-to-dehradun, Route: Chandigarh to Dehradun
```


jupyter
Red_Bus_Scrapping Last Checkpoint: 8 days ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)
[ ]:

```
print(CTU)

# Save the DataFrame to a CSV file
CTU.to_csv('CTU.csv', index=False)

# Close the browser
driver.quit()
```

```
650  https://www.redbus.in/bus-tickets/narnaul-to-c...
651  https://www.redbus.in/bus-tickets/chandigarh-t...

                  Bus Name                            Bus Type  \
0              HRTC - 6709            Ordinary 3+2 Non AC Seater
1              HRTC - 6639     Himsuta AC Seater Volvo/Scania 2+2
2               HRTC - 763            Ordinary 3+2 Non AC Seater
3               HRTC - 602            Ordinary 3+2 Non AC Seater
4               HRTC - 724            Ordinary 3+2 Non AC Seater
..                     ...                                   ...
647      Ram Dalal Holidays          Volvo A/C Semi Sleeper (2+2)
648          Laxmi holidays          Volvo A/C Semi Sleeper (2+2)
649        Bharmani Travels          Volvo A/C Semi Sleeper (2+2)
650  Hari Das Tour &Travels  Bharat Benz A/C Seater /Sleeper (2+1)
651               HRTC - 573            Ordinary 3+2 Non AC Seater

    Departure Time Duration Reaching Time Bus Rating    Price  \
0            21:40  05h 00m         02:40        3.9  INR 285
```