Question 4: Use Webscraping to Extract GME Revenue Data

Use the `requests` library to download the webpage https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/stock.html. Save the text of the response as a variable named `html_data_2`.

47]:

```
url1="https://cf-courses-data.s3.us.cloud-object-storage.appdomain.cloud/IBMDeveloperSkillsNetwork-PY0220EN-SkillsNetwork/labs/project/stock.html"
html_data_2  = requests.get(url1).text
print(html_data_2)
```

```
<!DOCTYPE html> •••
```

Parse the html data using `beautiful_soup` using parser i.e `html5lib` or `html.parser`.

48]:

```
soup1 = BeautifulSoup(html_data_2, "html5lib")
```

Using `BeautifulSoup` or the `read_html` function extract the table with `GameStop Revenue` and store it into a dataframe named `gme_revenue`. The dataframe should have columns `Date` and `Revenue`. Make sure the comma and dollar sign is removed from the `Revenue` column.

> **Note: Use the method similar to what you did in question 2.**

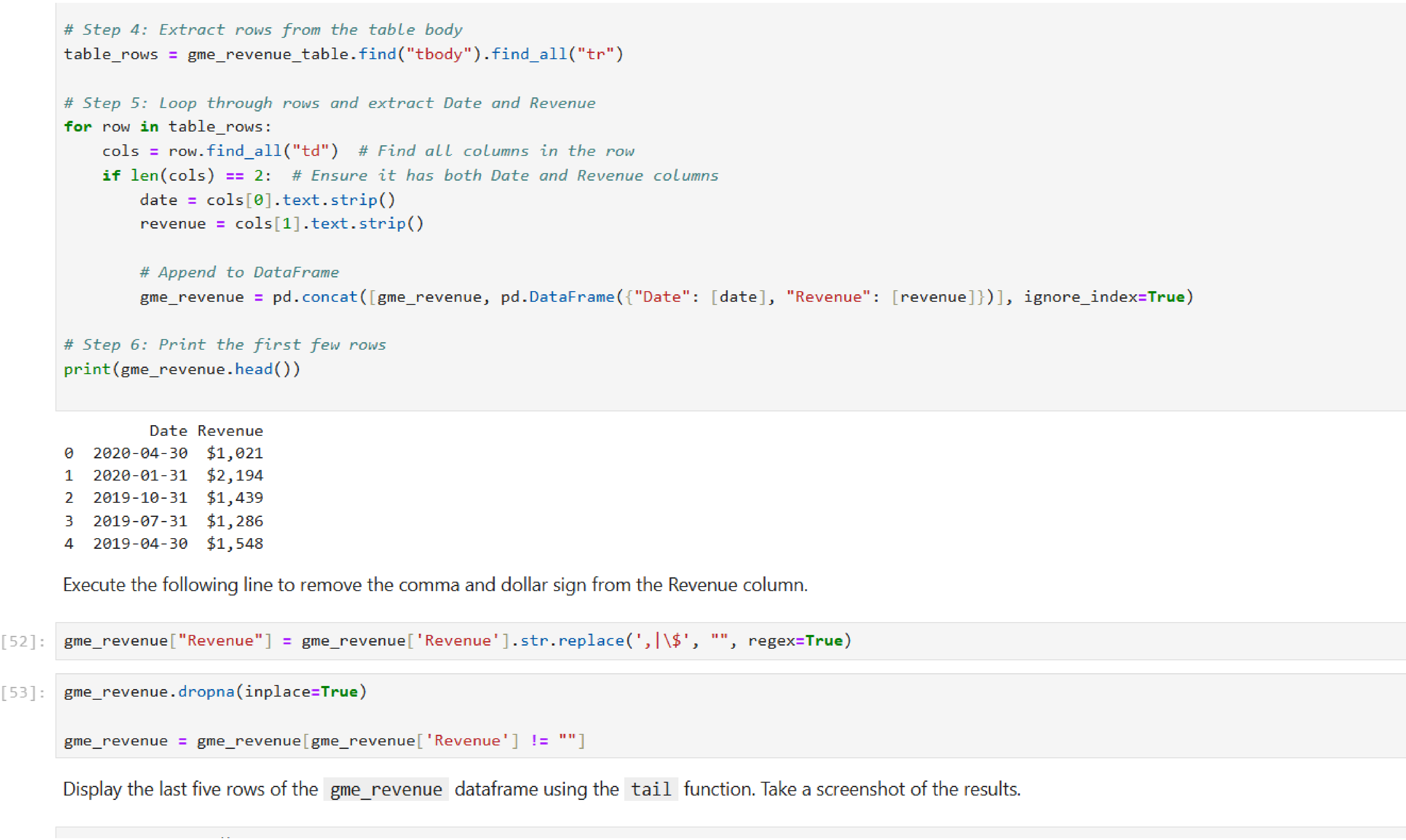
▶ Click here if you need help locating the table

49]:

```

# Step 2: Locate the correct table (Quarterly Revenue)
tables = soup1.find_all("table")  # Find all tables
gme_revenue_table = tables[1]  # The second table contains revenue data

# Step 3: Create an empty DataFrame
gme_revenue = pd.DataFrame(columns=["Date", "Revenue"])
```



```python
# Step 4: Extract rows from the table body
table_rows = gme_revenue_table.find("tbody").find_all("tr")

# Step 5: Loop through rows and extract Date and Revenue
for row in table_rows:
    cols = row.find_all("td")  # Find all columns in the row
    if len(cols) == 2:  # Ensure it has both Date and Revenue columns
        date = cols[0].text.strip()
        revenue = cols[1].text.strip()

        # Append to DataFrame
        gme_revenue = pd.concat([gme_revenue, pd.DataFrame({"Date": [date], "Revenue": [revenue]})], ignore_index=True)

# Step 6: Print the first few rows
print(gme_revenue.head())
```

```
         Date Revenue
0  2020-04-30  $1,021
1  2020-01-31  $2,194
2  2019-10-31  $1,439
3  2019-07-31  $1,286
4  2019-04-30  $1,548
```

Execute the following line to remove the comma and dollar sign from the Revenue column.

[52]:

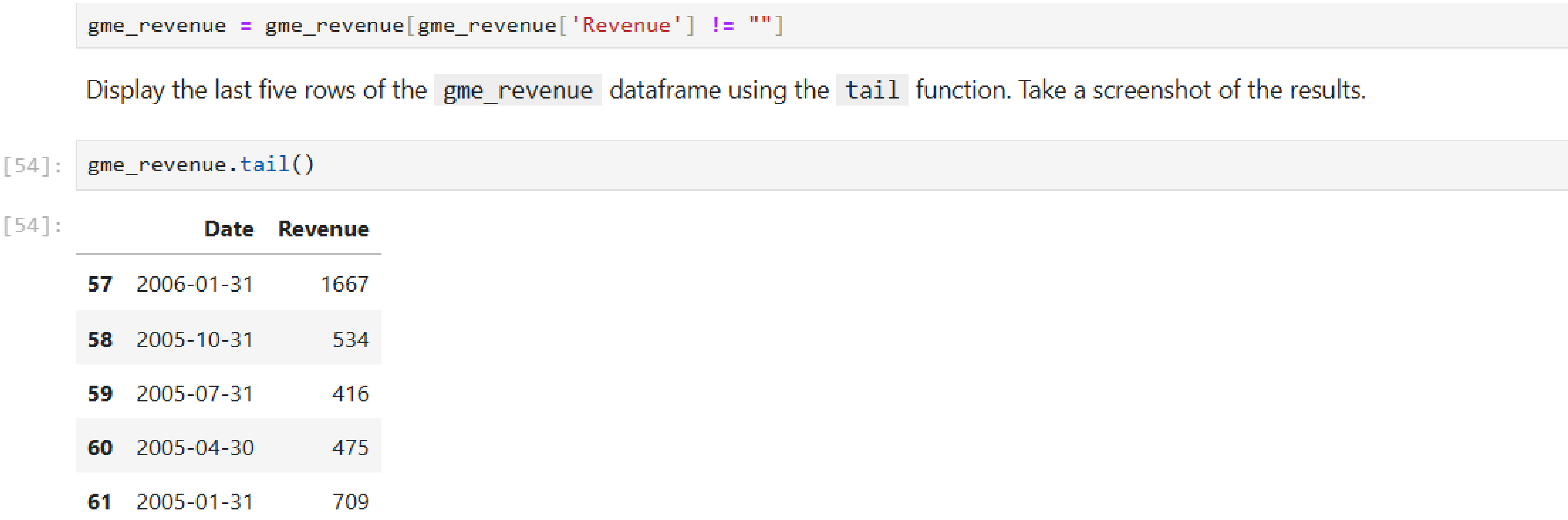
```python
gme_revenue["Revenue"] = gme_revenue['Revenue'].str.replace(',|\$', "", regex=True)
```

[53]:

```python
gme_revenue.dropna(inplace=True)

gme_revenue = gme_revenue[gme_revenue['Revenue'] != ""]
```

Display the last five rows of the `gme_revenue` dataframe using the `tail` function. Take a screenshot of the results.



```
gme_revenue = gme_revenue[gme_revenue['Revenue'] != ""]
```

Display the last five rows of the `gme_revenue` dataframe using the `tail` function. Take a screenshot of the results.

[54]:

```
gme_revenue.tail()
```

[54]:

	Date	Revenue
57	2006-01-31	1667
58	2005-10-31	534
59	2005-07-31	416
60	2005-04-30	475
61	2005-01-31	709