Web Scraping & Excel Export: Largest U.S. Companies by Revenue

This project scrapes data from Wikipedia on the largest companies in the United States by revenue. The data is parsed using BeautifulSoup and exported to Excel using Pandas. The script includes logging and error handling to simulate real-world API/data pipeline logic.

Fetch Webpage Content and Parse Table from HTML

Using `requests` to fetch the Wikipedia page. Added error handling and logging to ensure stability. Used `BeautifulSoup` to extract the main table from the page and convert it into a DataFrame.

In [275…
```
url = 'https://en.wikipedia.org/wiki/List_of_largest_companies_in_the_United_States

import requests
from bs4 import BeautifulSoup
import logging

logging.basicConfig(level=logging.INFO)

try:
    response = requests.get(url)
    response.raise_for_status()
    soup = BeautifulSoup(response.text, 'html.parser')
    logging.info("Successfully fetched and parsed the webpage.")
except requests.exceptions.RequestException as e:
    logging.error(f"Failed to fetch data: {e}")
    soup = None
```

```
INFO:root:Successfully fetched and parsed the webpage.
```

In [ ]:
```
print(soup)
```

In [ ]:
```
soup.find_all('table')[1] # there are two tables with the same class names.
```

In [ ]:
```
table = soup.find('table', class_ = 'wikitable sortable')
print(table)
```

In [279…
```
word_titles = table.find_all('th')
```

In [ ]:
```
print(word_titles)
```

In [ ]:
```
word_table_titles = [title.text.strip() for title in word_titles] # printing out th
print(word_table_titles)
```

In [282…
```
import pandas as pd
```

In [283…
```
df = pd.DataFrame(columns = word_table_titles)
print(df)
```

```
Empty DataFrame
Columns: [Rank, Name, Industry, Revenue (USD millions), Revenue growth, Employees, H
eadquarters]
Index: []
```

In [ ]:
```
column_data = table.find_all('tr')
print(column_data)
```

In [ ]:
```
for row in column_data[1:]: #  add[1:] cause there is null in first row. #W we are
    row_data= row.find_all('td') # gives out individaul data
    individual_row_data = [data.text.strip() for data in row_data]
    print(individual_row_data)
    length = len(df)
    df.loc[length] = individual_row_data
```

Export Data to Excel

The DataFrame is exported to `wikipedia_data.xlsx` using `pandas.to_excel()`.

In [286…
```
df.to_excel("wikipedia_data.xlsx", index=False, sheet_name='WikiData')
```

In [287…
```
df.head(10)
```

Out[287...

	Rank	Name	Industry	Revenue (USD millions)	Revenue growth	Employees	Headquarters
0	1	Walmart	Retail	648,125	6.0%	2,100,000	Bentonville, Arkansas
1	2	Amazon	Retail and cloud computing	574,785	11.9%	1,525,000	Seattle, Washington
2	3	Apple	Electronics industry	383,482	-2.8%	161,000	Cupertino, California
3	4	UnitedHealth Group	Healthcare	371,622	14.6%	440,000	Minnetonka, Minnesota
4	5	Berkshire Hathaway	Conglomerate	364,482	20.7%	396,500	Omaha, Nebraska
5	6	CVS Health	Healthcare	357,776	10.9%	259,500	Woonsocket, Rhode Island
6	7	ExxonMobil	Petroleum industry	344,582	-16.7%	61,500	Spring, Texas
7	8	Alphabet	Technology and cloud computing	307,394	8.7%	182,502	Mountain View, California
8	9	McKesson Corporation	Health	276,711	4.8%	48,000	Irving, Texas
9	10	Cencora	Pharmacy wholesale	262,173	9.9%	44,000	Conshohocken, Pennsylvania

Summary

This project simulates a real-world data pipeline by:

- Fetching data from the web
- Parsing structured information
- Exporting results into Excel
- Including logging and error handling

This mirrors how API-based workflows operate in data analytics.