Python Project for Data Engineering

There is a total of 10 points possible for the final project.

1. Upload the image ‘Task_1_log_function.png’. This should show the code for the function `log_progress()` used in the project. (1 point)

```
[30]: log_file = 'C:/Users/NARAYAN1.JHA/Downloads/log_file.txt'
def log_progress(message):
    timestamp_format = '%Y-%h-%d-%H:%M:%S' # Year-Monthname-Day-Hour-Minute-Second
    now = datetime.now() # get current timestamp
    timestamp = now.strftime(timestamp_format)
    with open(log_file,"a") as f:
        f.write(timestamp + ',' + message + '\n')
```

2. Upload the image ‘Task_2a_extract.png’. This should be the snapshot of the html code obtained by inspecting the table on the webpage. The contents of the first row should be expanded and visible. (1 point)

```
Elements  Console  Sources  Network  Performance  Memory  Application  Security  Lighthouse  Recorder  Performance insights
  <meta property="mw:PageProp/toc">
  ▶<h2>…</h2>
  ▶<p>…</p>
  ▶<style data-mw-deduplicate="TemplateStyles:r1096954695/mw-parser-output/.tmulti">…</style>
  ▶<div class="thumb tmulti tright">…</div>
  ▼<table class="wikitable sortable mw-collapsible jquery-tablesorter mw-made-collapsible">
    ▶<thead>…</thead>
    ▼<tbody>
      ▼<tr>
        <td>1 </td>
        ▼<td>
          ▶<span class="flagicon">…</span>
           <a href="/web/20230908091635/https://en.wikipedia.org/wiki/JPMorgan_Chase" title="JPMorgan Chase">JPMorgan Chase</a>
        </td>
        <td>432.92 </td>
      </tr> == $0
      ▼<tr>
        <td>2 </td>
        ▼<td>
          ▶<span class="flagicon">…</span>
           <a href="/web/20230908091635/https://en.wikipedia.org/wiki/Bank_of_America" title="Bank of America">Bank of America</a>
        </td>
        <td>231.52 </td>
      </tr>
```

3. Upload the image ‘Task_2b_extract.png’. This should show the code for the function ‘extract()’ used in the project. (1 point)

```
[11]: def extract(url):
    # Read the tables from the URL using pandas
    try:
        tables = pd.read_html(url)
        if tables:
            # Assuming the first table matches the criteria
            df = tables[1]
            print(df)
            return df
        else:
            print("No tables found with the given attributes.")
            return pd.DataFrame()  # Return an empty DataFrame
    except Exception as e:
        print(f"An error occurred: {e}")
        return pd.DataFrame()
```

4. Upload the image ‘Task_2c_extract.png’. This should be the output obtained by executing the function call. (1 point)

```
[17]: df=extract(URL)

   Rank                              Bank name  Market cap (US$ billion)
0     1                           JPMorgan Chase                    432.92
1     2                          Bank of America                    231.52
2     3  Industrial and Commercial Bank of China                    194.56
3     4                Agricultural Bank of China                    160.68
4     5                                 HDFC Bank                    157.91
5     6                               Wells Fargo                    155.87
6     7                         HSBC Holdings PLC                    148.90
7     8                             Morgan Stanley                    140.83
8     9                   China Construction Bank                    139.82
9    10                            Bank of China                    136.81
```

5. Upload the image ‘Task_3a_transform.png’. This should show the code for the function ‘transform’ used in the project. (1 point)

```
[85]: def transform(df):
          df = df.rename(columns = {'Bank name' : 'Name'})
          df = df.rename(columns = {'Market cap (US$ billion)' : 'MC_USD_Billion'})
          df['MC_GBP_Billion'] = np.round(df['MC_USD_Billion']*0.8,2)
          df['MC_EUR_Billion'] = np.round(df['MC_USD_Billion']*0.93,2)
          df['MC_INR_Billion'] = np.round(df['MC_USD_Billion']*82.95,2)
          return df
```

6. Upload the image ‘Task_3b_transform.png’. This should be the output of the final transformed dataframe. (1 point)

```
[98]: df=transform(df)
      print(df)

   Rank                                Name  MC_USD_Billion  \
0     1                           JPMorgan Chase          432.92
1     2                          Bank of America          231.52
2     3  Industrial and Commercial Bank of China          194.56
3     4                Agricultural Bank of China          160.68
4     5                                 HDFC Bank          157.91
5     6                               Wells Fargo          155.87
6     7                         HSBC Holdings PLC          148.90
7     8                             Morgan Stanley          140.83
8     9                   China Construction Bank          139.82
9    10                            Bank of China          136.81

   MC_GBP_Billion  MC_EUR_Billion  MC_INR_Billion
0          346.34          402.62        35910.71
1          185.22          215.31        19204.58
2          155.65          180.94        16138.75
3          128.54          149.43        13328.41
4          126.33          146.86        13098.63
5          124.70          144.96        12929.42
6          119.12          138.48        12351.26
7          112.66          130.97        11681.85
8          111.86          130.03        11598.07
9          109.45          127.23        11348.39
```

7. Upload the image ‘Task_4_CSV.png’. This should be the contents of the CSV file created from the final table. (1 point)

	A	B	C	D	E	F	G	H	I
1		Rank	Name	MC_USD_Billion	MC_GBP_Billion	MC_EUR_Billion	MC_INR_Billion		
2	0	1	JPMorgan Chase	432.92	346.34	402.62	35910.71		
3	1	2	Bank of America	231.52	185.22	215.31	19204.58		
4	2	3	Industrial and Commercial Bank of China	194.56	155.65	180.94	16138.75		
5	3	4	Agricultural Bank of China	160.68	128.54	149.43	13328.41		
6	4	5	HDFC Bank	157.91	126.33	146.86	13098.63		
7	5	6	Wells Fargo	155.87	124.7	144.96	12929.42		
8	6	7	HSBC Holdings PLC	148.9	119.12	138.48	12351.26		
9	7	8	Morgan Stanley	140.83	112.66	130.97	11681.85		
10	8	9	China Construction Bank	139.82	111.86	130.03	11598.07		
11	9	10	Bank of China	136.81	109.45	127.23	11348.39		
12									
13									

8. Upload the image ‘Task_4_5_save_file.png’. This should show the code for both `load_to_csv()` and `load_to_db()` functions used in the project. (1 point)

```
csv_file='C:/Users/NARAYAN1.JHA/Downloads/Largest_banks_data.csv'
def load_to_csv(csv_file,df):
    df.to_csv(csv_file)

def load_to_db(transformed_data, db_url, table_name, if_exists="replace"):
    try:
        # Create a database connection engine
        engine = create_engine(db_url)

        # Load DataFrame to SQL table
        df.to_sql(name=table_name, con=engine, if_exists=if_exists, index=False)

        print(f"Data successfully loaded into the '{table_name}' table.")
    except Exception as e:
        print(f"An error occurred while loading data to the database: {e}")
```

9. Upload the image ‘Task_6_SQL.png’. This should be the output of the SQL queries run on the database table. (1 point)

```
[95]: run_queries(table_name)

SELECT * FROM Bank_data
(1, 'JPMorgan Chase', 432.92, 346.34, 402.62, 35910.71)
(2, 'Bank of America', 231.52, 185.22, 215.31, 19204.58)
(3, 'Industrial and Commercial Bank of China', 194.56, 155.65, 180.94, 16138.75)
(4, 'Agricultural Bank of China', 160.68, 128.54, 149.43, 13328.41)
(5, 'HDFC Bank', 157.91, 126.33, 146.86, 13098.63)
(6, 'Wells Fargo', 155.87, 124.7, 144.96, 12929.42)
(7, 'HSBC Holdings PLC', 148.9, 119.12, 138.48, 12351.26)
(8, 'Morgan Stanley', 140.83, 112.66, 130.97, 11681.85)
(9, 'China Construction Bank', 139.82, 111.86, 130.03, 11598.07)
(10, 'Bank of China', 136.81, 109.45, 127.23, 11348.39)


SELECT AVG(MC_GBP_Billion) FROM Bank_data
(151.987,)


SELECT Name FROM Bank_data LIMIT 5
('JPMorgan Chase',)
('Bank of America',)
('Industrial and Commercial Bank of China',)
('Agricultural Bank of China',)
('HDFC Bank',)

[ ]:
```

10. Upload the image ‘Task_7_log_content.png’. This should be the contents of the log file ‘code_log.txt’. (1 point)

log_file - Notepad

File Edit View

```
2025-Jan-25-20:15:30,ETL Job Started
2025-Jan-25-20:15:30,Extract phase Started
Log Timestamp: 2025-Jan-25-20:15:35
 Rank                              Bank name  Market cap (US$ billion)
    1                          JPMorgan Chase                    432.92
    2                        Bank of America                    231.52
    3 Industrial and Commercial Bank of China                    194.56
    4              Agricultural Bank of China                    160.68
    5                               HDFC Bank                    157.91
    6                             Wells Fargo                    155.87
    7                       HSBC Holdings PLC                    148.90
    8                          Morgan Stanley                    140.83
    9                 China Construction Bank                    139.82
   10                           Bank of China                    136.81

2025-Jan-25-20:15:35,Extract phase Ended
2025-Jan-25-20:15:41,Transform phase Started
Log Timestamp: 2025-Jan-25-20:15:41
 Rank                                   Name  MC_USD_Billion  MC_GBP_Billion  MC_EUR_Billion  MC_INR_Billion
    1                          JPMorgan Chase          432.92          346.34          402.62        35910.71
    2                        Bank of America          231.52          185.22          215.31        19204.58
    3 Industrial and Commercial Bank of China          194.56          155.65          180.94        16138.75
    4              Agricultural Bank of China          160.68          128.54          149.43        13328.41
    5                               HDFC Bank          157.91          126.33          146.86        13098.63
    6                             Wells Fargo          155.87          124.70          144.96        12929.42
    7                       HSBC Holdings PLC          148.90          119.12          138.48        12351.26
    8                          Morgan Stanley          140.83          112.66          130.97        11681.85
    9                 China Construction Bank          139.82          111.86          130.03        11598.07
   10                           Bank of China          136.81          109.45          127.23        11348.39

2025-Jan-25-20:15:41,Transform phase Ended
2025-Jan-25-20:16:13,Load phase Started
2025-Jan-25-20:16:13,Loading data to csv
```

Ln 33, Col 41 100% Windows (C