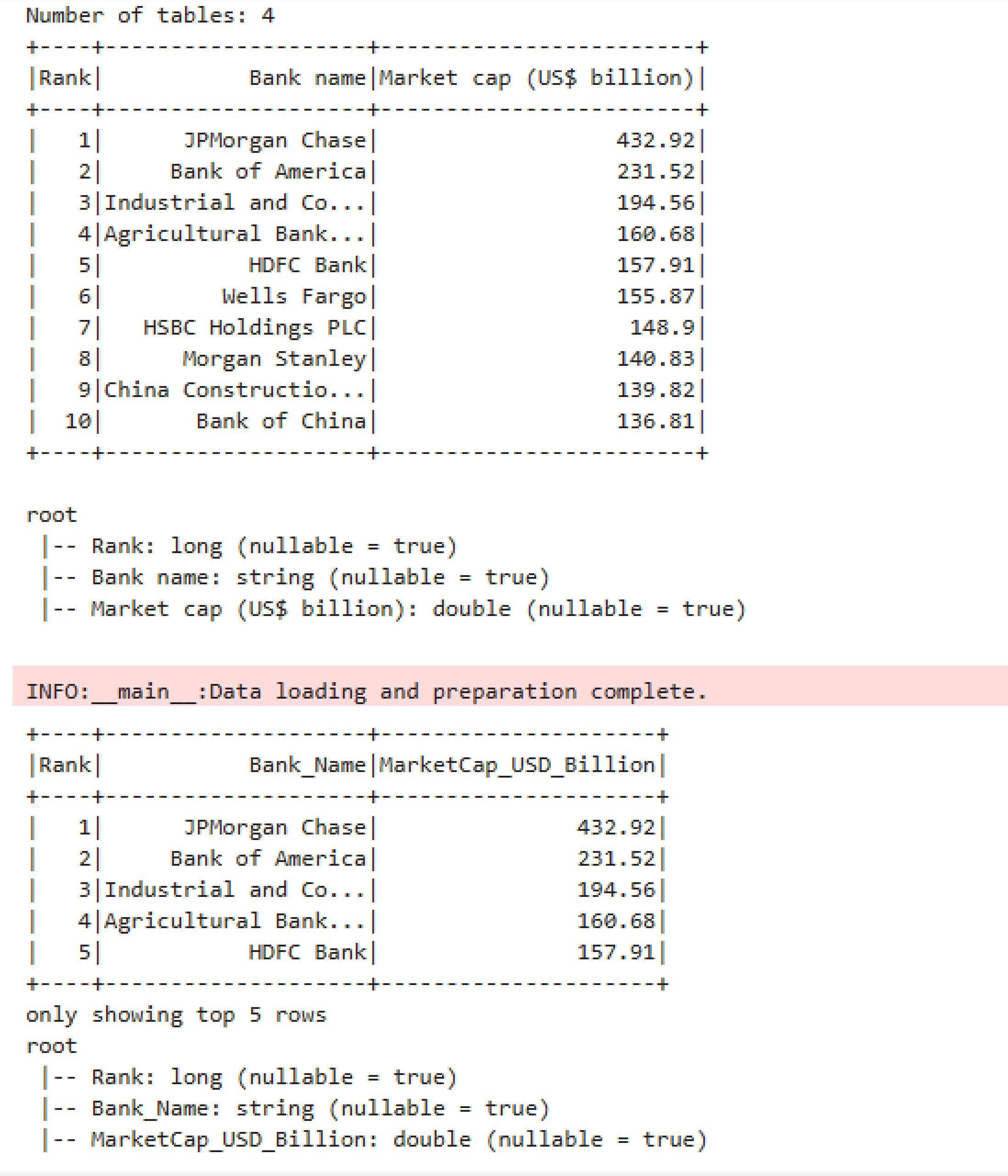
Report: Bank Data Analysis

1. Data Preparation

The dataset was successfully loaded into PySpark, containing 4 tables.
Columns were cleaned and standardized (Bank Name, MarketCap_USD_Billion).
The top banks by market capitalization include JPMorgan Chase, Bank of America, and HDFC Bank.



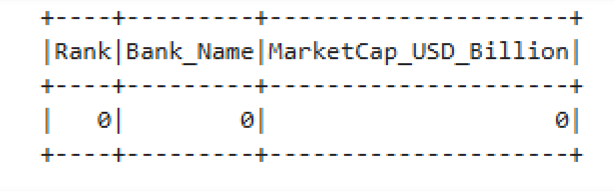
```
Number of tables: 4
+----+--------------------+------------------------+
|Rank|           Bank name|Market cap (US$ billion)|
+----+--------------------+------------------------+
|   1|      JPMorgan Chase|                  432.92|
|   2|     Bank of America|                  231.52|
|   3|Industrial and Co...|                  194.56|
|   4|Agricultural Bank...|                  160.68|
|   5|           HDFC Bank|                  157.91|
|   6|         Wells Fargo|                  155.87|
|   7|   HSBC Holdings PLC|                   148.9|
|   8|      Morgan Stanley|                  140.83|
|   9|China Constructio...|                  139.82|
|  10|       Bank of China|                  136.81|
+----+--------------------+------------------------+

root
 |-- Rank: long (nullable = true)
 |-- Bank name: string (nullable = true)
 |-- Market cap (US$ billion): double (nullable = true)

INFO:__main__:Data loading and preparation complete.
+----+--------------------+---------------------+
|Rank|           Bank_Name|MarketCap_USD_Billion|
+----+--------------------+---------------------+
|   1|      JPMorgan Chase|               432.92|
|   2|     Bank of America|               231.52|
|   3|Industrial and Co...|               194.56|
|   4|Agricultural Bank...|               160.68|
|   5|           HDFC Bank|               157.91|
+----+--------------------+---------------------+
only showing top 5 rows
root
 |-- Rank: long (nullable = true)
 |-- Bank_Name: string (nullable = true)
 |-- MarketCap_USD_Billion: double (nullable = true)
```

2. Data Cleaning

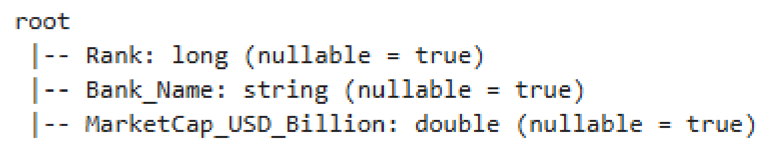
2.1 Handle Missing Values

Before cleaning, the schema showed valid column structures.
Rows with missing or invalid values (e.g., Rank = 0, Bank Name = 0, Market Cap = 0) were

identified and removed to ensure data consistency.



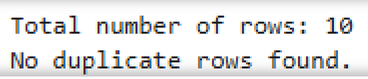
```
+----+---------+--------------------+
|Rank|Bank_Name|MarketCap_USD_Billion|
+----+---------+--------------------+
|   0|        0|                   0|
+----+---------+--------------------+
```



```
root
 |-- Rank: long (nullable = true)
 |-- Bank_Name: string (nullable = true)
 |-- MarketCap_USD_Billion: double (nullable = true)
```

2.2 Fixing Columns

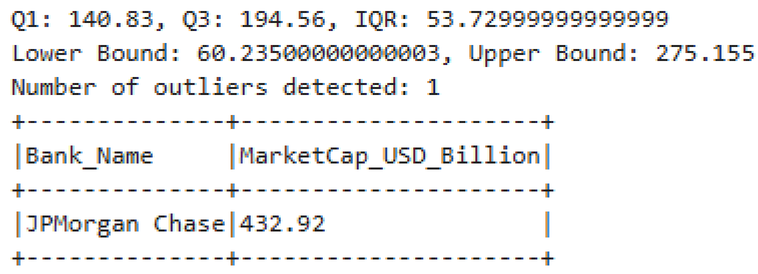
The MarketCap_USD_Billion column was already in numeric format, so no conversion was required.
The dataset contains 10 rows in total, and no duplicate records were found.



```
Total number of rows: 10
No duplicate rows found.
```

2.3 Handle Outliers

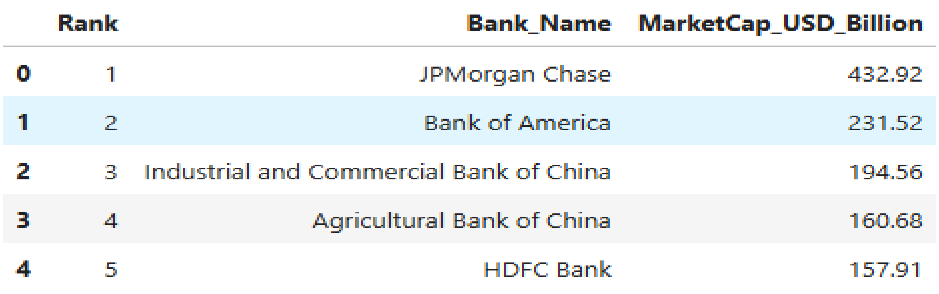
Outlier detection using the IQR method found one bank exceeding the upper bound.
JPMorgan Chase was identified as an outlier with a market capitalization of **432.92 USD billion**.



```
Q1: 140.83, Q3: 194.56, IQR: 53.72999999999999
Lower Bound: 60.23500000000003, Upper Bound: 275.155
Number of outliers detected: 1
+--------------+--------------------+
|Bank_Name     |MarketCap_USD_Billion|
+--------------+--------------------+
|JPMorgan Chase|432.92              |
+--------------+--------------------+
```

3. Exploratory Data Analysis
3.1 Conversion from PySpark to Pandas Data Frame

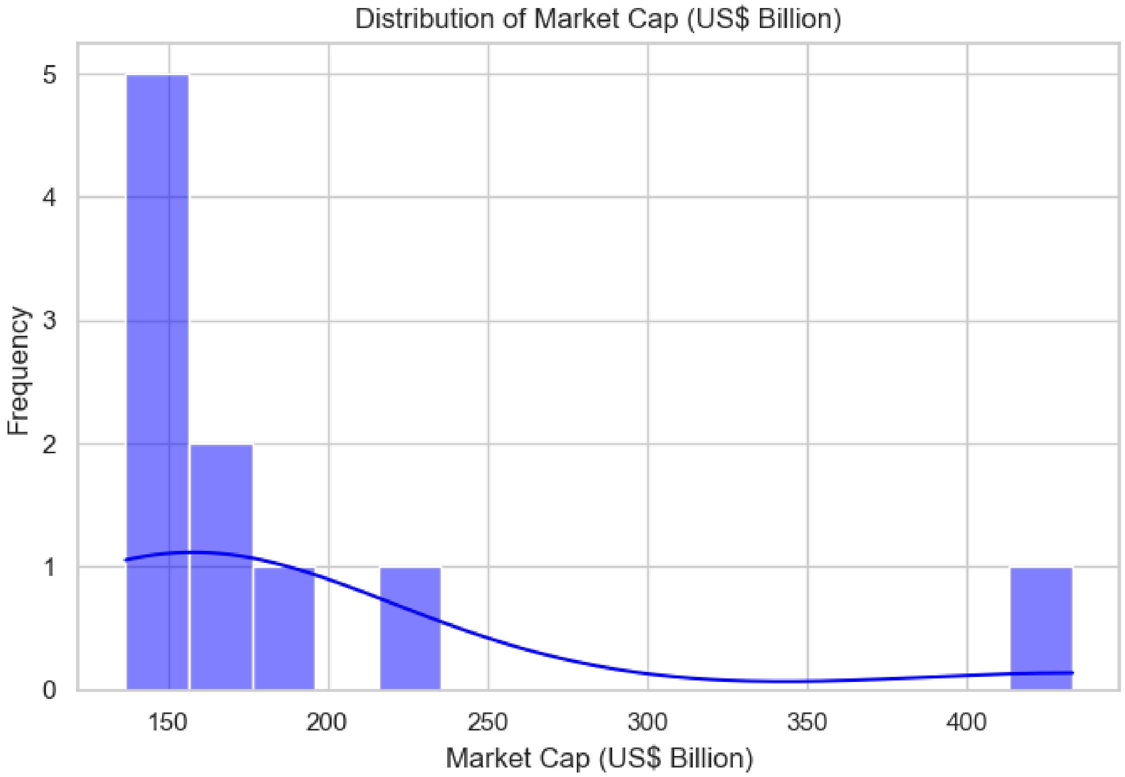
For visualization purposes, the PySpark Data Frame was converted into a Pandas Data Frame.
The converted dataset retains the ranking, bank names, and market capitalization values.
The top entries confirm that **JPMorgan Chase** holds the highest market cap, followed by **Bank of America** and the **Industrial and Commercial Bank of China**.



	Rank	Bank_Name	MarketCap_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

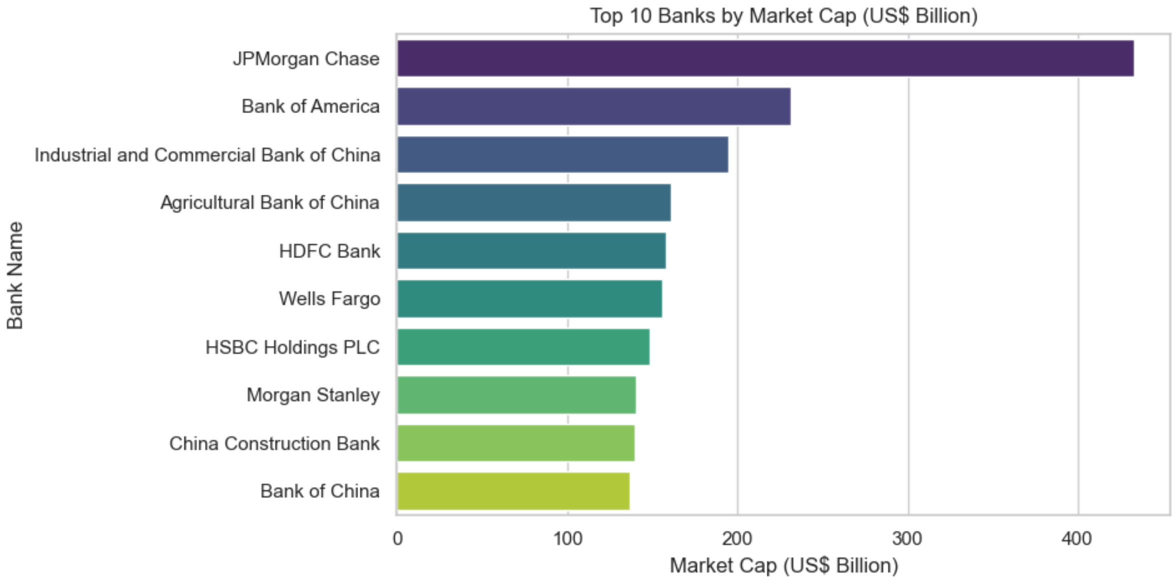
3.2 Market Capitalization

A histogram was plotted to analyze the distribution of market capitalization among the top banks.
The chart shows that most banks fall within the **150–400 USD billion** range, while **JPMorgan Chase** stands out with a significantly higher market cap.



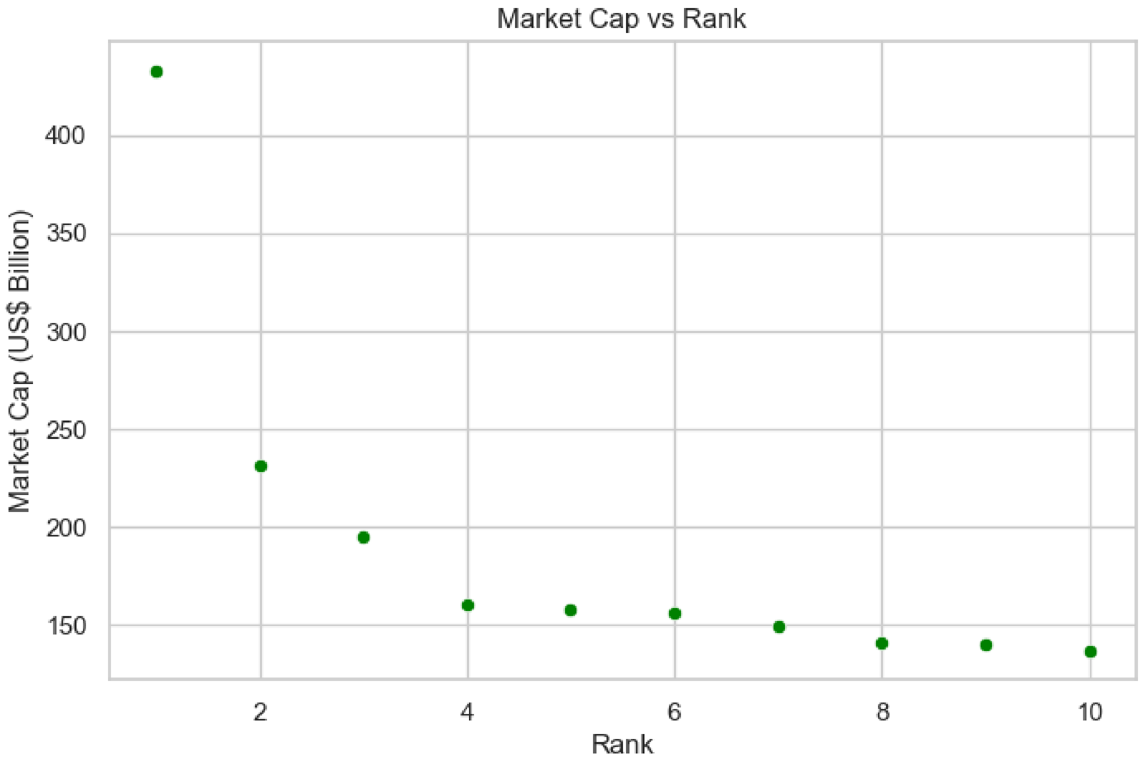
3.3 Top 10 Banks

A bar chart was created to visualize the top 10 banks by market capitalization.
The analysis shows **JPMorgan Chase** leading with the highest value (~450 USD billion), followed by **Bank of America** (~250 USD billion) and the **Industrial and Commercial Bank of China** (~180 USD billion).



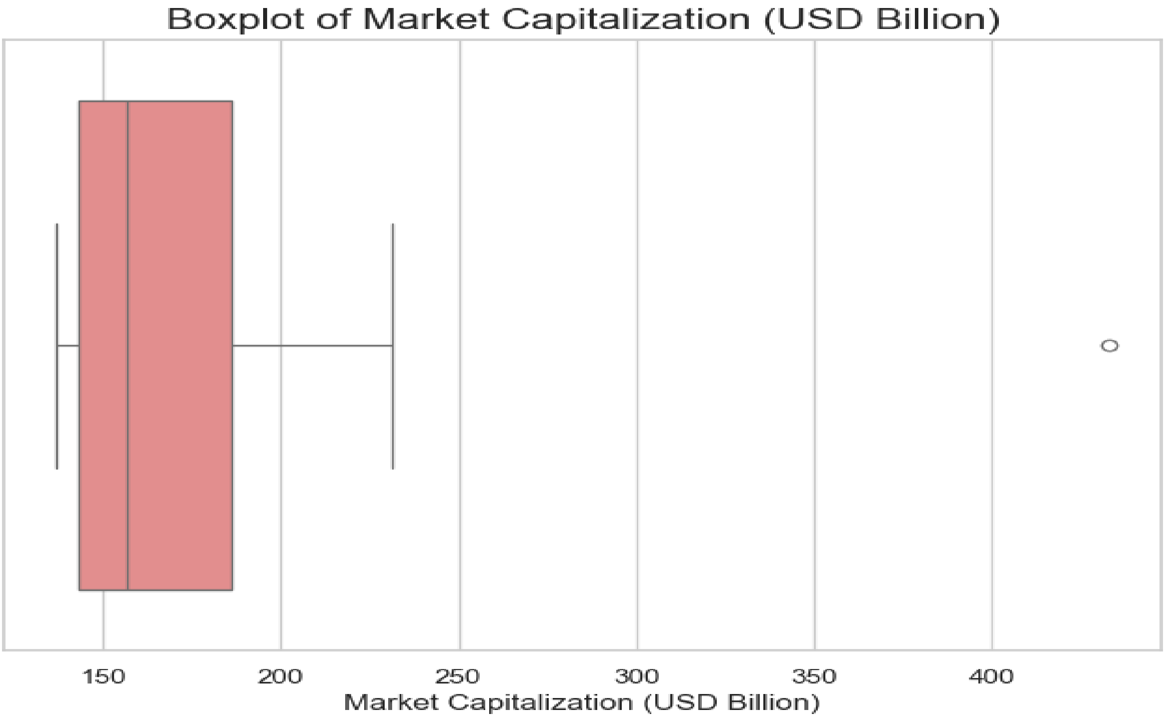
3.4 Market Cap vs Bank Ranking

A scatter plot was generated to visualize the relationship between bank ranking and market capitalization.
The chart shows an inverse trend — banks with higher rankings (lower rank numbers) generally have larger market capitalizations, with **JPMorgan Chase** standing out as the top performer.



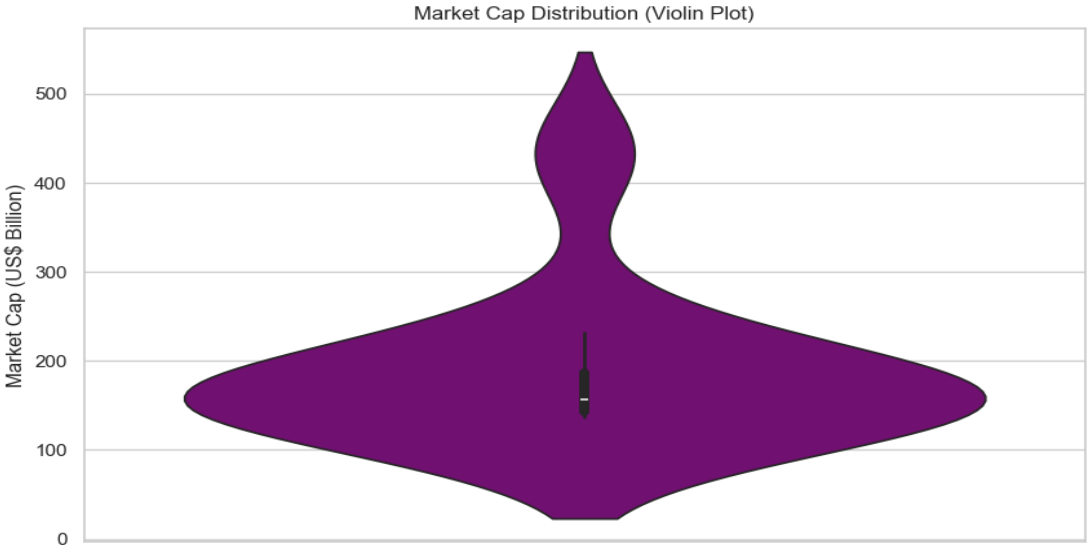
3.5 Market Cap Analysis

The boxplot shows that most banks have market capitalizations clustered between **140–200 USD billion**.
One significant outlier exists on the higher end, highlighting a bank with much larger capitalization compared to others.



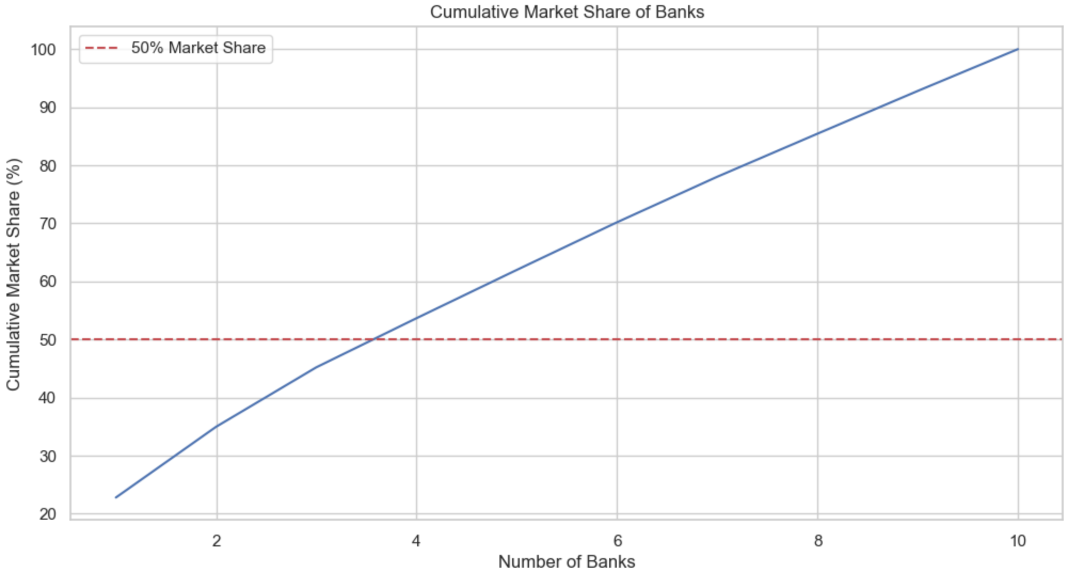
3.6 Market Cap Quartile Distribution

A violin plot was generated to visualize the quartile distribution of bank market capitalizations. Most banks are concentrated in the **150–200 USD billion** range, while a few banks lie at the upper extreme, creating a wider spread.



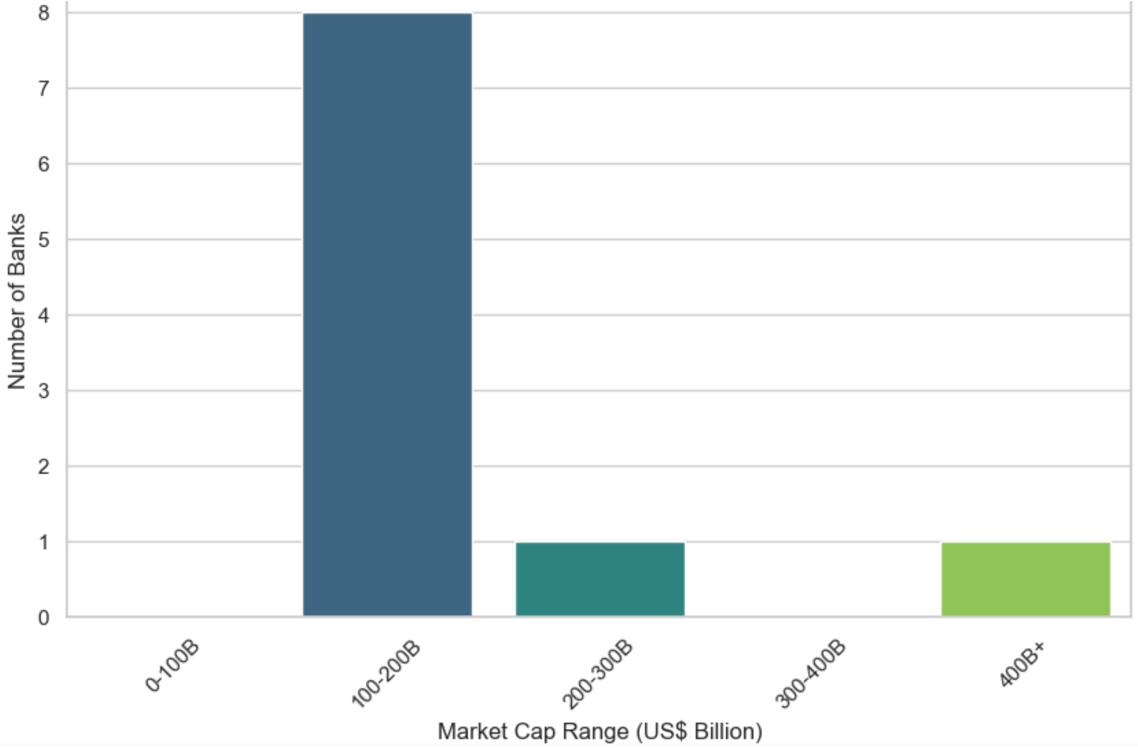
3.7 Cumulative Market Share Analysis

A line plot was created to analyze the cumulative market share of the top banks.
The results show that a small number of leading banks account for nearly **50% of the total market capitalization**, indicating a highly concentrated market structure.



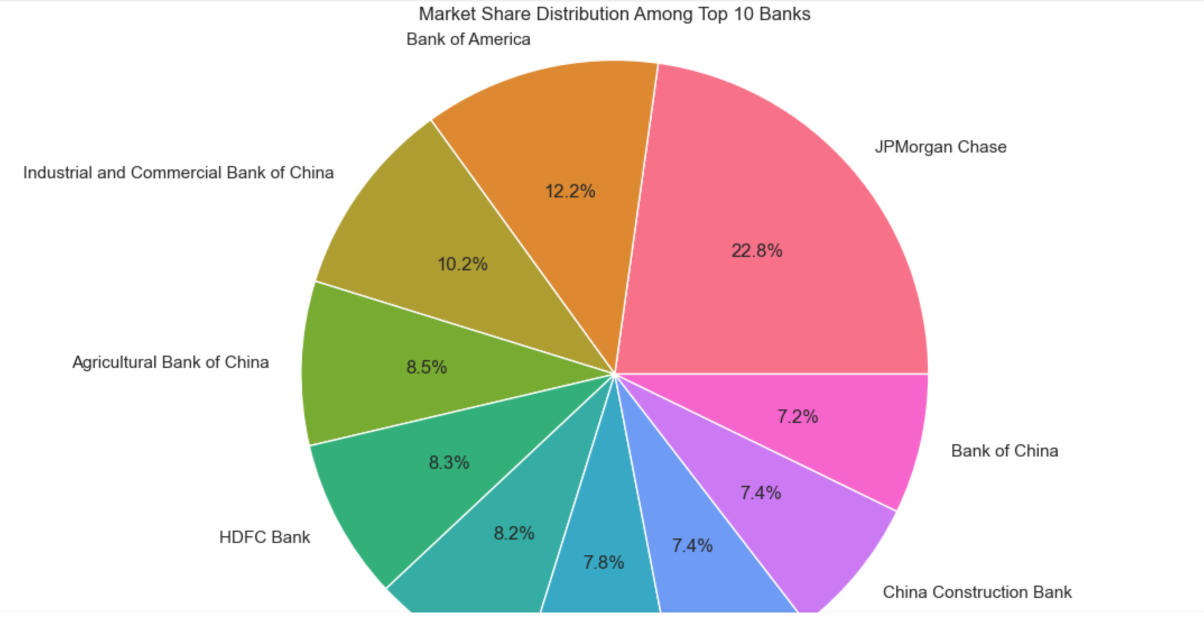
3.8 Categorising Banks

Banks were grouped into market capitalization ranges and visualized using a bar chart.
The distribution shows that **most banks (8 out of 10)** fall in the **100–200B USD** range, while only one bank lies above **400B USD**, highlighting market dominance by a single major player.



3.9 Visualise Market Share Distribution

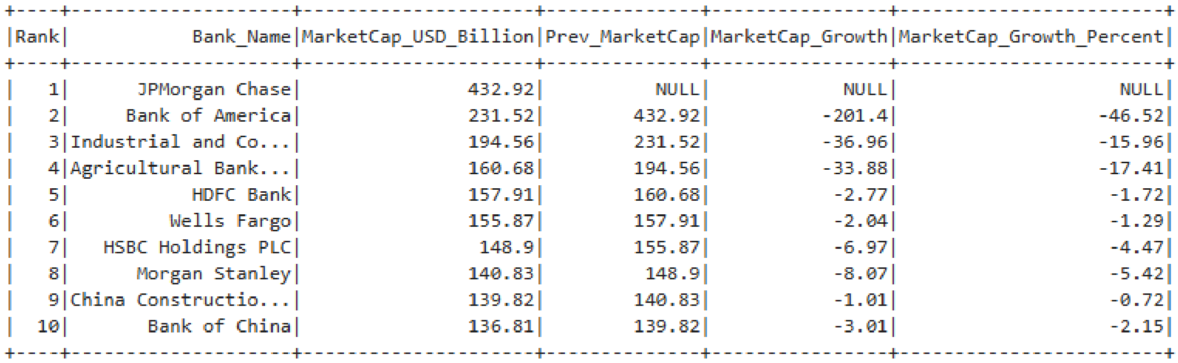
A pie chart was created to display the market share of the top 10 banks.
JPMorgan Chase dominates with **22.8%**, followed by **Bank of America (12.2%)** and the **Industrial and Commercial Bank of China (10.2%)**.
The remaining banks each contribute between **7–8%**, showing a moderately balanced distribution apart from the clear leader.



4. ETL and Querying

4.1 Market Capitalization Analysis

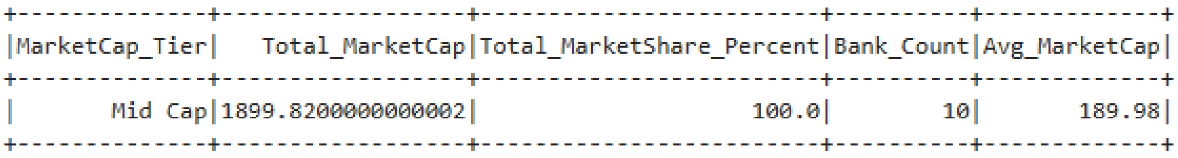
Advanced market capitalization analysis was performed with growth metrics.
The results show that **JPMorgan Chase** leads with the highest market cap (432.92B USD), while subsequent banks reflect declining values compared to the previous entry.
The steepest drop is seen for **Bank of America (-46.52%)**, followed by notable declines across other banks, highlighting a skewed distribution of growth.



```
+----+--------------------+---------------------+--------------+----------------+------------------------+
|Rank|           Bank_Name|MarketCap_USD_Billion|Prev_MarketCap|MarketCap_Growth|MarketCap_Growth_Percent|
+----+--------------------+---------------------+--------------+----------------+------------------------+
|   1|      JPMorgan Chase|               432.92|          NULL|            NULL|                    NULL|
|   2|     Bank of America|               231.52|        432.92|          -201.4|                  -46.52|
|   3|Industrial and Co...|               194.56|        231.52|          -36.96|                  -15.96|
|   4|Agricultural Bank...|               160.68|        194.56|          -33.88|                  -17.41|
|   5|           HDFC Bank|               157.91|        160.68|           -2.77|                   -1.72|
|   6|         Wells Fargo|               155.87|        157.91|           -2.04|                   -1.29|
|   7|   HSBC Holdings PLC|                148.9|        155.87|           -6.97|                   -4.47|
|   8|      Morgan Stanley|               140.83|         148.9|           -8.07|                   -5.42|
|   9|China Constructio...|               139.82|        140.83|           -1.01|                   -0.72|
|  10|       Bank of China|               136.81|        139.82|           -3.01|                   -2.15|
+----+--------------------+---------------------+--------------+----------------+------------------------+
```

4.2 Market Concentration Analysis

Banks were categorized into market capitalization tiers to assess concentration.
All 10 banks fall under the **Mid Cap** tier, contributing a combined market cap of **~1899.82B USD** (100% share), with an average of **189.98B USD** per bank.

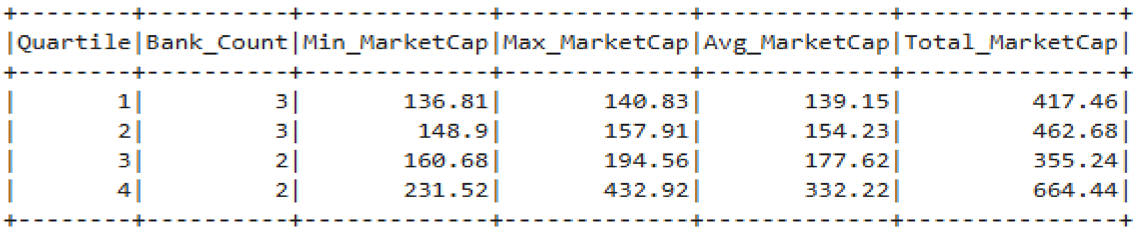


```
+--------------+------------------+-------------------------+----------+-------------+
|MarketCap_Tier|   Total_MarketCap|Total_MarketShare_Percent|Bank_Count|Avg_MarketCap|
+--------------+------------------+-------------------------+----------+-------------+
|       Mid Cap|1899.8200000000002|                    100.0|        10|       189.98|
+--------------+------------------+-------------------------+----------+-------------+
```

4.3 Market Capitalization Distribution

Quartile analysis was performed to study the distribution of bank market capitalization.

- Q1 and Q2 contain the majority of banks (6 in total) with market caps between **136B–158B USD**.
- Q3 includes 2 banks averaging **177.62B USD**.
- Q4 highlights 2 banks with the highest values, averaging **332.22B USD**, dominated by JPMorgan Chase.

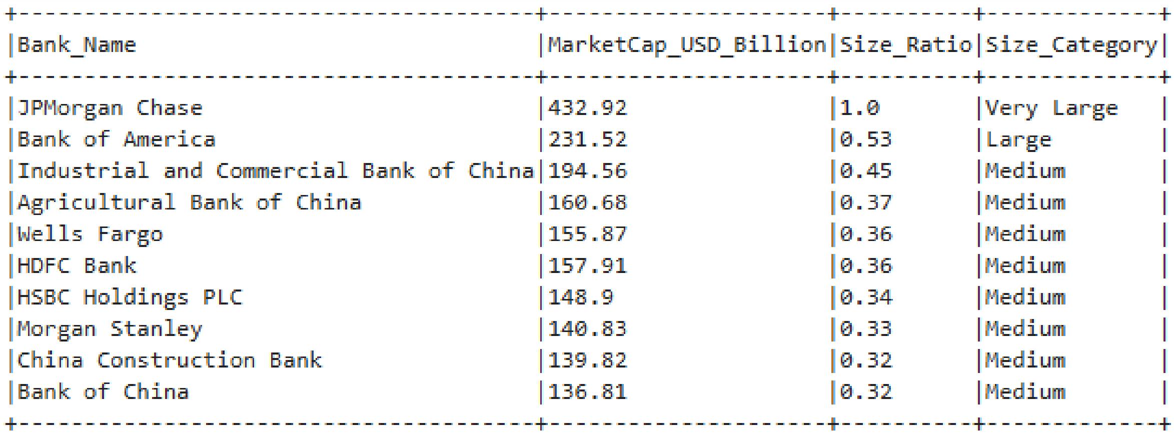


```
+--------+----------+-------------+-------------+-------------+---------------+
|Quartile|Bank_Count|Min_MarketCap|Max_MarketCap|Avg_MarketCap|Total_MarketCap|
+--------+----------+-------------+-------------+-------------+---------------+
|       1|         3|       136.81|       140.83|       139.15|         417.46|
|       2|         3|        148.9|       157.91|       154.23|         462.68|
|       3|         2|       160.68|       194.56|       177.62|         355.24|
|       4|         2|       231.52|       432.92|       332.22|         664.44|
+--------+----------+-------------+-------------+-------------+---------------+
```

4.4 Comparative Size Analysis

Banks were classified by relative market size using size ratios.

- **JPMorgan Chase** stands out as the only **Very Large** bank with a ratio of 1.0.
- **Bank of America** falls in the **Large** category (0.53).
- All other banks are categorized as **Medium**, with size ratios between 0.32–0.45.

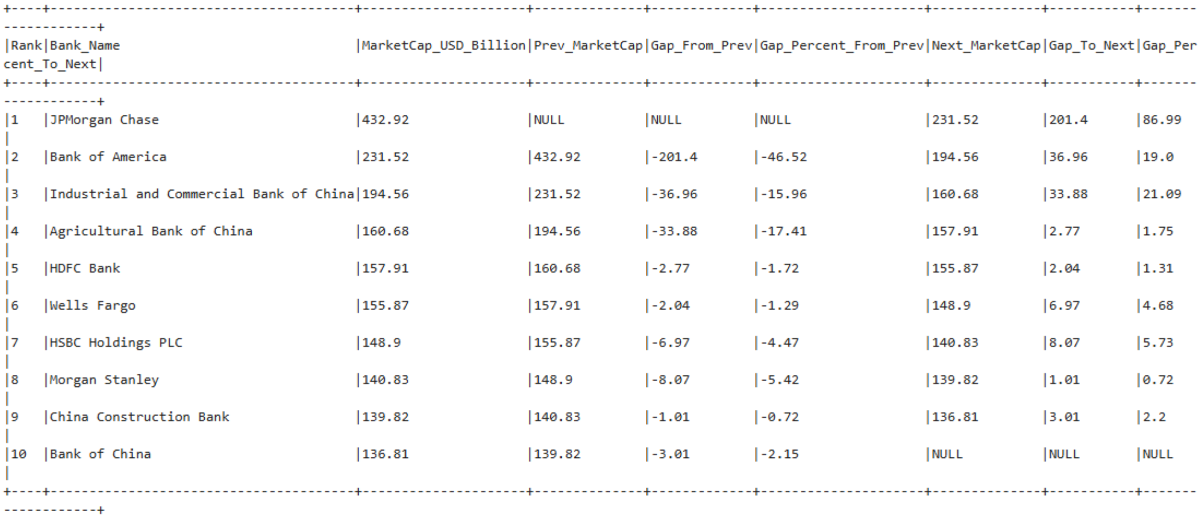


```
+---------------------------------------+---------------------+----------+-------------+
|Bank_Name                              |MarketCap_USD_Billion|Size_Ratio|Size_Category|
+---------------------------------------+---------------------+----------+-------------+
|JPMorgan Chase                         |432.92               |1.0       |Very Large   |
|Bank of America                        |231.52               |0.53      |Large        |
|Industrial and Commercial Bank of China|194.56               |0.45      |Medium       |
|Agricultural Bank of China             |160.68               |0.37      |Medium       |
|Wells Fargo                            |155.87               |0.36      |Medium       |
|HDFC Bank                              |157.91               |0.36      |Medium       |
|HSBC Holdings PLC                      |148.9                |0.34      |Medium       |
|Morgan Stanley                         |140.83               |0.33      |Medium       |
|China Construction Bank                |139.82               |0.32      |Medium       |
|Bank of China                          |136.81               |0.32      |Medium       |
+---------------------------------------+---------------------+----------+-------------+
```

4.5 Market Growth Analysis

Market growth gaps were analyzed between consecutive banks.
The largest gap is between **JPMorgan Chase** and **Bank of America**, with a difference of

201.4B USD (86.99%).
Banks ranked 4th to 10th show relatively smaller gaps, indicating a more evenly distributed market capitalization among them.

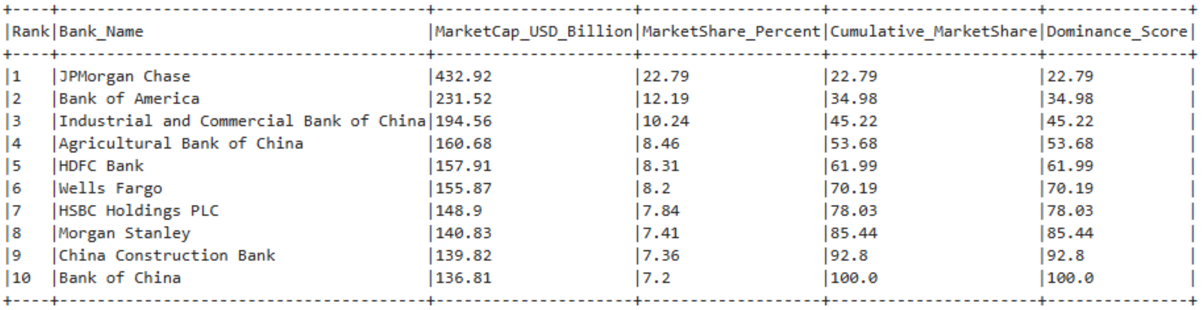


```
+----+---------------------------------------+---------------------+--------------+-------------+---------------------+--------------+-----------+-------
------------+
|Rank|Bank_Name                              |MarketCap_USD_Billion|Prev_MarketCap|Gap_From_Prev|Gap_Percent_From_Prev|Next_MarketCap|Gap_To_Next|Gap_Per
cent_To_Next|
+----+---------------------------------------+---------------------+--------------+-------------+---------------------+--------------+-----------+-------
------------+
|1   |JPMorgan Chase                         |432.92               |NULL          |NULL         |NULL                 |231.52        |201.4      |86.99
|
|2   |Bank of America                        |231.52               |432.92        |-201.4       |-46.52               |194.56        |36.96      |19.0
|
|3   |Industrial and Commercial Bank of China|194.56               |231.52        |-36.96       |-15.96               |160.68        |33.88      |21.09
|
|4   |Agricultural Bank of China             |160.68               |194.56        |-33.88       |-17.41               |157.91        |2.77       |1.75
|
|5   |HDFC Bank                              |157.91               |160.68        |-2.77        |-1.72                |155.87        |2.04       |1.31
|
|6   |Wells Fargo                            |155.87               |157.91        |-2.04        |-1.29                |148.9         |6.97       |4.68
|
|7   |HSBC Holdings PLC                      |148.9                |155.87        |-6.97        |-4.47                |140.83        |8.07       |5.73
|
|8   |Morgan Stanley                         |140.83               |148.9         |-8.07        |-5.42                |139.82        |1.01       |0.72
|
|9   |China Construction Bank                |139.82               |140.83        |-1.01        |-0.72                |136.81        |3.01       |2.2
|
|10  |Bank of China                          |136.81               |139.82        |-3.01        |-2.15                |NULL          |NULL       |NULL
|
+----+---------------------------------------+---------------------+--------------+-------------+---------------------+--------------+-----------+-------
------------+
```

4.6. Market Dominance Analysis

Market dominance was evaluated based on individual and cumulative market shares.

- **JPMorgan Chase** leads with **22.79%** share alone.
- Top 3 banks cumulatively hold **45.22%**, showing significant dominance.
- Market share declines steadily, reaching **100%** at Rank 10.
- Dominance Score mirrors cumulative share, indicating each bank’s weight in total dominance.



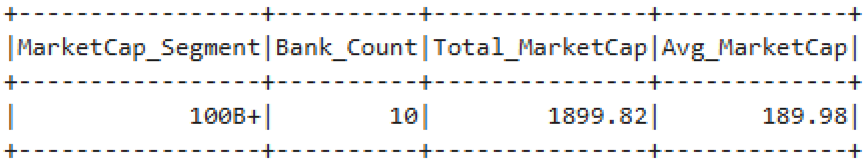
```
+----+---------------------------------------+---------------------+-------------------+----------------------+---------------+
|Rank|Bank_Name                              |MarketCap_USD_Billion|MarketShare_Percent|Cumulative_MarketShare|Dominance_Score|
+----+---------------------------------------+---------------------+-------------------+----------------------+---------------+
|1   |JPMorgan Chase                         |432.92               |22.79              |22.79                 |22.79          |
|2   |Bank of America                        |231.52               |12.19              |34.98                 |34.98          |
|3   |Industrial and Commercial Bank of China|194.56               |10.24              |45.22                 |45.22          |
|4   |Agricultural Bank of China             |160.68               |8.46               |53.68                 |53.68          |
|5   |HDFC Bank                              |157.91               |8.31               |61.99                 |61.99          |
|6   |Wells Fargo                            |155.87               |8.2                |70.19                 |70.19          |
|7   |HSBC Holdings PLC                      |148.9                |7.84               |78.03                 |78.03          |
|8   |Morgan Stanley                         |140.83               |7.41               |85.44                 |85.44          |
|9   |China Construction Bank                |139.82               |7.36               |92.8                  |92.8           |
|10  |Bank of China                          |136.81               |7.2                |100.0                 |100.0          |
+----+---------------------------------------+---------------------+-------------------+----------------------+---------------+
```

4.7. Segment-Wise Bank Analysis

Banks were grouped based on market capitalization segments.

- All **10 banks** fall in the **100B+ USD** segment.
- Combined market cap: **1899.82B USD**
- Average per bank: **189.98B USD**

  This indicates a **high-cap segment concentration** among top global banks.

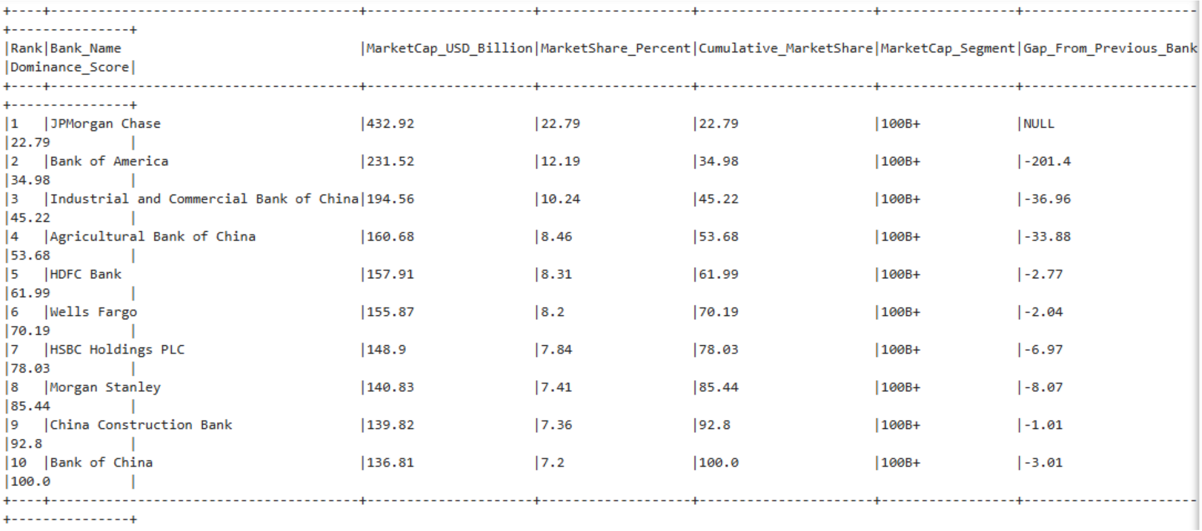


```
+-----------------+----------+---------------+-------------+
|MarketCap_Segment|Bank_Count|Total_MarketCap|Avg_MarketCap|
+-----------------+----------+---------------+-------------+
|            100B+|        10|        1899.82|       189.98|
+-----------------+----------+---------------+-------------+
```

4.8. Performance Dashboard

A consolidated dashboard presents key performance indicators across the top 10 banks.

- All banks fall under the **100B+ market cap** segment.
- **JPMorgan Chase** leads with the highest market cap (**432.92B USD**) and **dominance score of 22.79**.
- **Gap from previous bank** highlights significant drop after Rank 1 (−201.4B USD).
- **Cumulative market share** reaches 100% by Rank 10.



```
+----+---------------------------------------+---------------------+-------------------+----------------------+-----------------+----------------------
+---------------+
|Rank|Bank_Name                              |MarketCap_USD_Billion|MarketShare_Percent|Cumulative_MarketShare|MarketCap_Segment|Gap_From_Previous_Bank
|Dominance_Score|
+----+---------------------------------------+---------------------+-------------------+----------------------+-----------------+----------------------
+---------------+
|1   |JPMorgan Chase                         |432.92               |22.79              |22.79                 |100B+            |NULL
|22.79          |
|2   |Bank of America                        |231.52               |12.19              |34.98                 |100B+            |-201.4
|34.98          |
|3   |Industrial and Commercial Bank of China|194.56               |10.24              |45.22                 |100B+            |-36.96
|45.22          |
|4   |Agricultural Bank of China             |160.68               |8.46               |53.68                 |100B+            |-33.88
|53.68          |
|5   |HDFC Bank                              |157.91               |8.31               |61.99                 |100B+            |-2.77
|61.99          |
|6   |Wells Fargo                            |155.87               |8.2                |70.19                 |100B+            |-2.04
|70.19          |
|7   |HSBC Holdings PLC                      |148.9                |7.84               |78.03                 |100B+            |-6.97
|78.03          |
|8   |Morgan Stanley                         |140.83               |7.41               |85.44                 |100B+            |-8.07
|85.44          |
|9   |China Construction Bank                |139.82               |7.36               |92.8                  |100B+            |-1.01
|92.8           |
|10  |Bank of China                          |136.81               |7.2                |100.0                 |100B+            |-3.01
|100.0          |
+----+---------------------------------------+---------------------+-------------------+----------------------+-----------------+----------------------
+---------------+
```