Report: Banking Data Analysis

When working with Big Data, it is important to understand that the datasets are often too large to fit into the memory of a single machine. Throughout this project I will be leveraging PySpark a Python-based interface which is built on top of Apache Spark. PySpark enables distributed computing and integrates well with Hadoop Distributed File System (HDFS) for easy access and retrieval of information of large datasets.

1. Data Preparation / Data Extraction

The primary goal of Data Preparation is to extract the data and make it suitable for further analysis. The data is extracted from Wikipedia. Our focus is centered towards the top 10 largest banks in the world and analyzing their market capitalization. Market capitalization plays an important role in our analysis, since it helps determine the company's size and its worth in comparison to its peers.

1.1. Initializing a SparkSession

The first and most important step of any Big Data processing task with PySpark is initializing a SparkSession. It is the unified entry point for all Spark functionalities; without it, it's not possible to access or leverage any functionality that PySpark offers.

A SparkSession is created and verified to ensure that the session is currently active, allowing us to leverage Spark's functionalities.

1.2. Logging in PySpark

I created a helper function designed to help me log messages for debugging, informational, warning, error and critical messages during the runtime of my notebook. My logging function is integrated with "py4j" which serves as a bridge between Python and the Java Virtual Machine (JVM), allowing efficient tracking of logs for better troubleshooting.

My logging function has been well integrated in my ETL workflow and will generate a detailed log file with *"different levels of information"*.

1.3. Extraction of Banking Dataset (Via Web Scraping)

The table required for my analysis was extracted from Wikipedia. The information is from 8th September 2023.

Pandas library was used to perform the extraction of this table using the method *`pd.read_html`*. This method is widely used to read HTML tables from a webpage and returns a list of DataFrames.

1.4. Converting to PySpark DataFrame

The extracted Banking Dataset was subsequently converted to PySpark DataFrame to leverage Spark's scalability and efficient access to Big Data.

1. **Feature Renaming :** The features were standardized as per PySpark naming conventions.
2. **Schema Analysis :** The data types of all the features were analyzed thoroughly. The overall schema was analyzed as well to maintain Data Integrity.

1.5. Extracting the Exchange Rate Dataset

The Exchange Rate Dataset was provided in Excel format. After analyzing the exchange rate dataset, I made the decision to load the Dataset via Pandas and subsequently convert it into a PySpark DataFrame. Schema Analysis was performed after conversion to ensure that the data types were suitable for further processing.

Note: *PySpark DataFrames were cached from time-to-time to ensure fast access.*

2. Data Cleaning / Data Transformation

In this section, a range of transformations were performed to make the data suitable for further analysis. Some of those transformations include handling of missing values, correcting columns where necessary, feature engineering and handling of outliers.

2.1. Standardizing Columns

The Bank Name feature was standardized to ensure that the feature doesn't have any leading or trailing white spaces, thus preventing data inconsistency issues.

2.2. Handling Missing Values

Checking for missing values is as important as handling missing values. A sanity check was performed. Our Dataset didn't have a single missing value.

Despite our Dataset not having any missing values, for scalability purpose I have dropped rows with missing values in Bank Name and Market Cap feature.

2.2.1. Global Rank Imputation (Focused on Scalability)

Missing Values in the Global Rank feature were handled using a specialized approach. The Dataset was ordered in descending order on the basis of Market Capitalization and row numbers were then assigned to impute the Global Rank feature.

2.3. Fixing Columns and Feature Engineering

2.3.1. Data Integrity Check

The data types of all our features were accurate. My analysis revealed that the total number of rows we were currently dealing with was 10 Rows. Each row referencing a single bank.

2.3.2. Duplicate Record Handling (Bank Name)

A specialized approach was taken to ensure that duplicate entries in the bank name feature was handled efficiently.

This was accomplished by :

1. Creating a temporary column where bank names were converted to lower case and stripped of leading / trailing white spaces.
2. Subsequently dropping duplicate records based on the standardized column.

2.3.3. Feature Engineering

A scalable currency dictionary was created from the PySpark Exchange Rate Dataset. The dictionary currently consists of 3 currencies and their respective exchange rate.

The scalable currency dictionary was leveraged to derive new features in the main Banking Dataset. The derived features were the Market Capitalization of Banks in different currencies. The derived features were renamed as per Spark's naming conventions and to maintain consistency.

2.4. Handling Outliers

A detailed outlier analysis was performed using an efficient and scalable method. This was achieved by leveraging PySpark functionality to convert the PySpark DataFrame into Panda's on PySpark DataFrame *(via **pandas_api** method)* which still leverages the power's of Spark's distributed processing and parallel processing and gives us the same performance while working with Big Data.

An Outlier Prediction function was defined to leverage built-in pandas methods such as ***quantiles()***. The function was designed in a way to use the widely known statistical measure that is Interquartile Range (IQR).

The IQR method is perfect because it measures the dispersion or spread of the data and is less affected by extreme outliers. Our user-defined function revealed a potential outlier JPMorgan Chase. However, further analysis revealed that JPMorgan isn't an outlier and is actually a market leader.

2.5. Loading the Data

This is the final stage of the ETL pipeline. In this stage, the transformed data was stored in a local folder stimulating an S3 Object Storage bucket, mirroring a typical ETL on cloud workflow. The data was stored in Parquet format, which is widely used for storing Big Data, making it ready for reuse and further analysis.

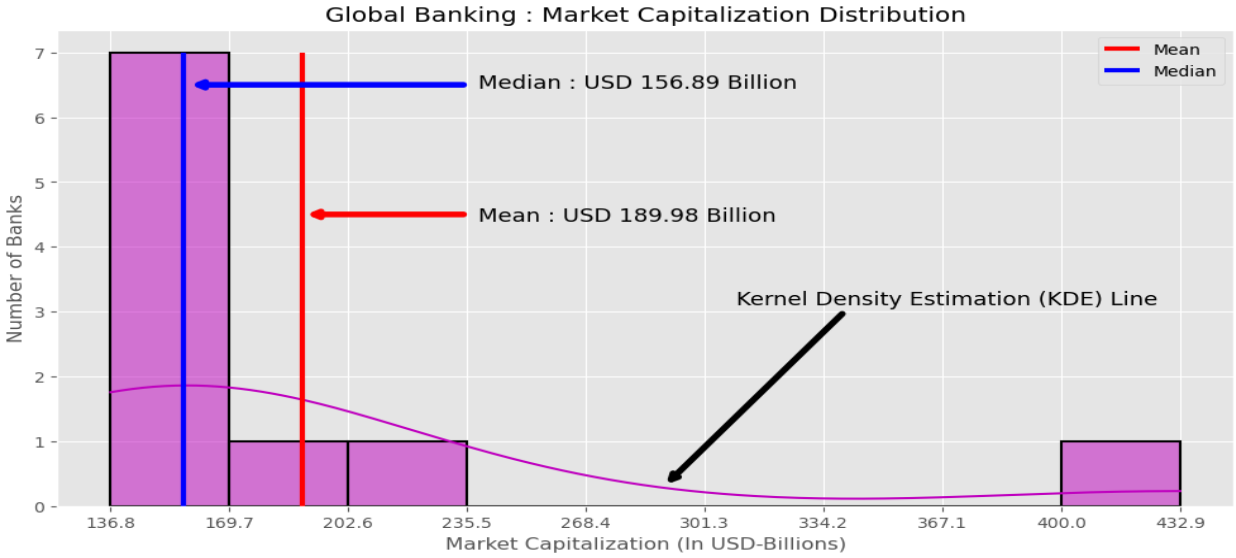
3. Exploratory Data Analysis: Finding Patterns and Trends

3.1. Convert PySpark DataFrame to Pandas DataFrame for visualization.

In this section, steps were taken to ensure that the data was ready for visualization. A specialized approach was used, accounting for scalability and necessary precautions when working with Big Data. The approach involved a conditional check on the number of rows in the PySpark DataFrame. If the count was less than 1 lakh the entire dataset was converted to Pandas DataFrame directly. Otherwise, 10% of the data was sampled before conversion.

3.2. Market Capitalization Distribution Analysis (Histogram)

The Majority of the banks are clustered at the lower end of the market capitalization scale, specifically lying between USD 136.8 Billion to USD 235.5 Billion. Our analysis suggests that 9 out of 10 banks lie between this range.



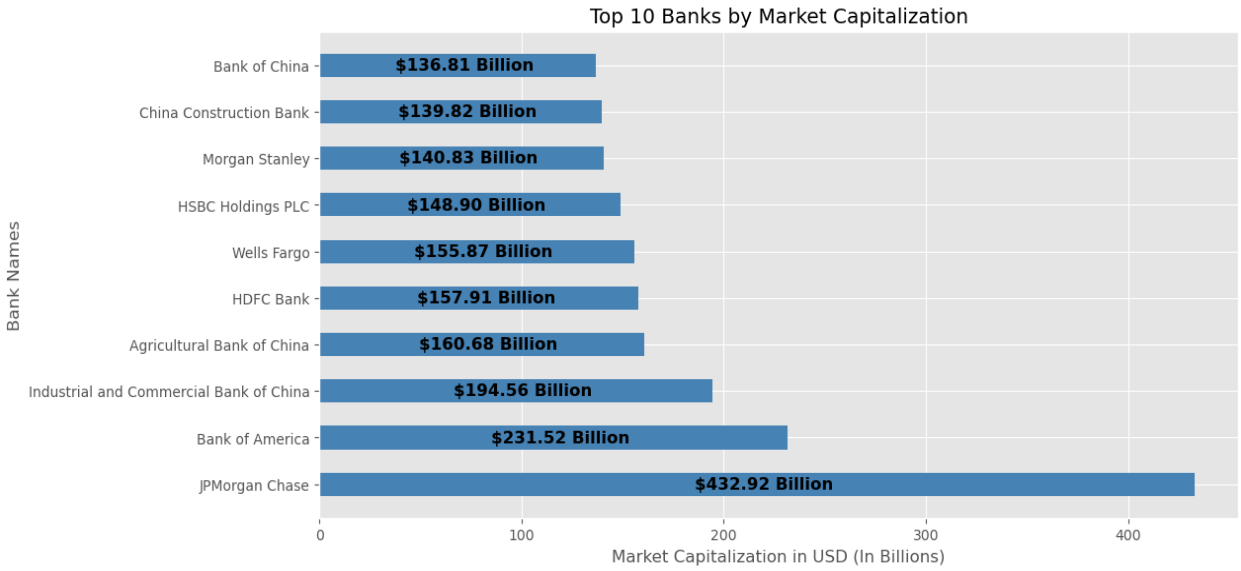
The Mean is USD 189.98 Billion and the Median is USD 156.89 Billion. This discrepancy is due to the fact that mean is heavily impacted by extreme values, whereas median is a more stable measure of central tendency. It also indicates that the data is positively skewed, this skewness is driven by a small number of extremely large banks or market leaders which pull the mean value upward.

The Kernel Density Estimation (KDE) line accounts for this skewness of the data and indicates the presence of extreme values in our dataset.

3.3. Top 10 banks by market capitalization using a Bar Chart.

The Market Capitalization feature was studied using a bar chart, since histogram isn't used to analyze the feature in depth, it is rather used to analyse the distribution of a feature.

The Top 10 Banks with the highest market capitalization were identified and plotted using a Bar Chart. As per the Top 10 Banks (Ranked on basis of Market Capitalization), Bank of China had the lowest Market Capitalization (USD 136.81 Billion) whereas JP Morgan Chase had the highest Market Capitalization (USD 432.92 Billion).



JPMorgan Chase's capitalization is nearly double of that of the second most dominant bank as per our analysis which is Bank of America (USD 231.52 Billion). There is a strong geographical mix of Chinese and US Banking institutions dominating the Top 10 list.

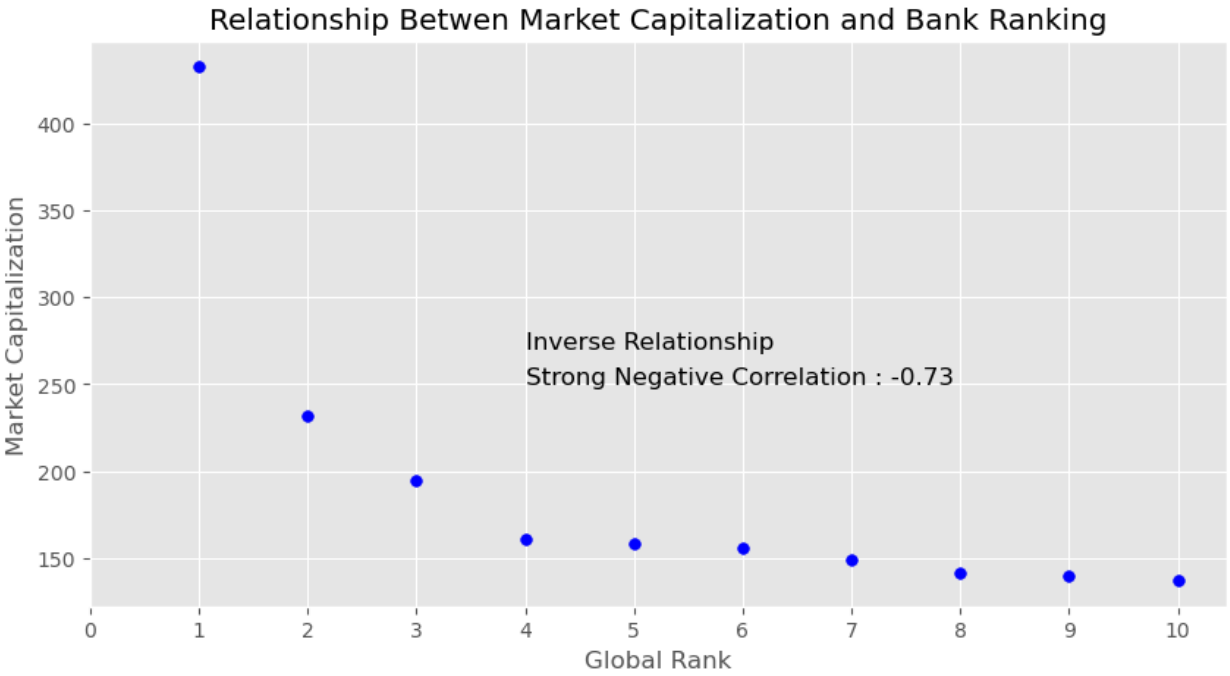
- **Chinese Banking Institutions –** Bank of China, China Construction Bank, Agricultural Bank of Chinda, Industrial and Commercial Bank of China.

- **United States Banking Institutions –** JPMorgan Chase, Bank of America, Wells Fargo, Morgan Stanley.

- **Other Banks –** HSBC Holdings PLC from United Kingdom and HDFC Bank from India.

3.4. Market capitalization VS Bank Ranking (Study of Relationship between the two features)

The first step towards studying the relationship between the two features was a correlation analysis. A strong negative correlation was found between the two features, meaning if one increases the other decreases, this is also known as an inverse relationship.



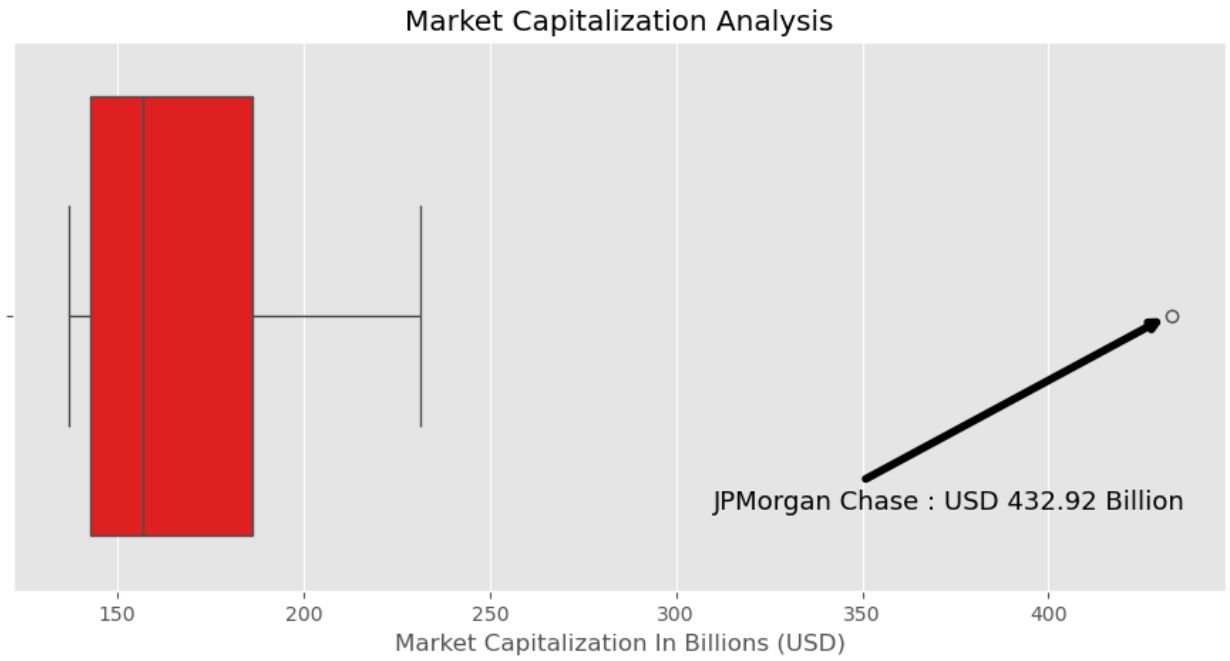
The Scatter Plot shows a clear inverse relationship between Market Capitalization and Bank Ranking (Global Rank). As the Bank Ranking increases Market Capitalization decreases and vice versa. This means that Market Capitalization would be the highest when the Bank Ranking is 1.

3.5. Market Capitalization Analysis (Spread and Outliers)

The main focus of this analysis was to figure out the spread of the Market Capitalization feature and study the outliers. Our analysis revealed that 90% of the data in this feature lies within the Interquartile Range (USD 140.8 Billion to USD 194.56 Billion).

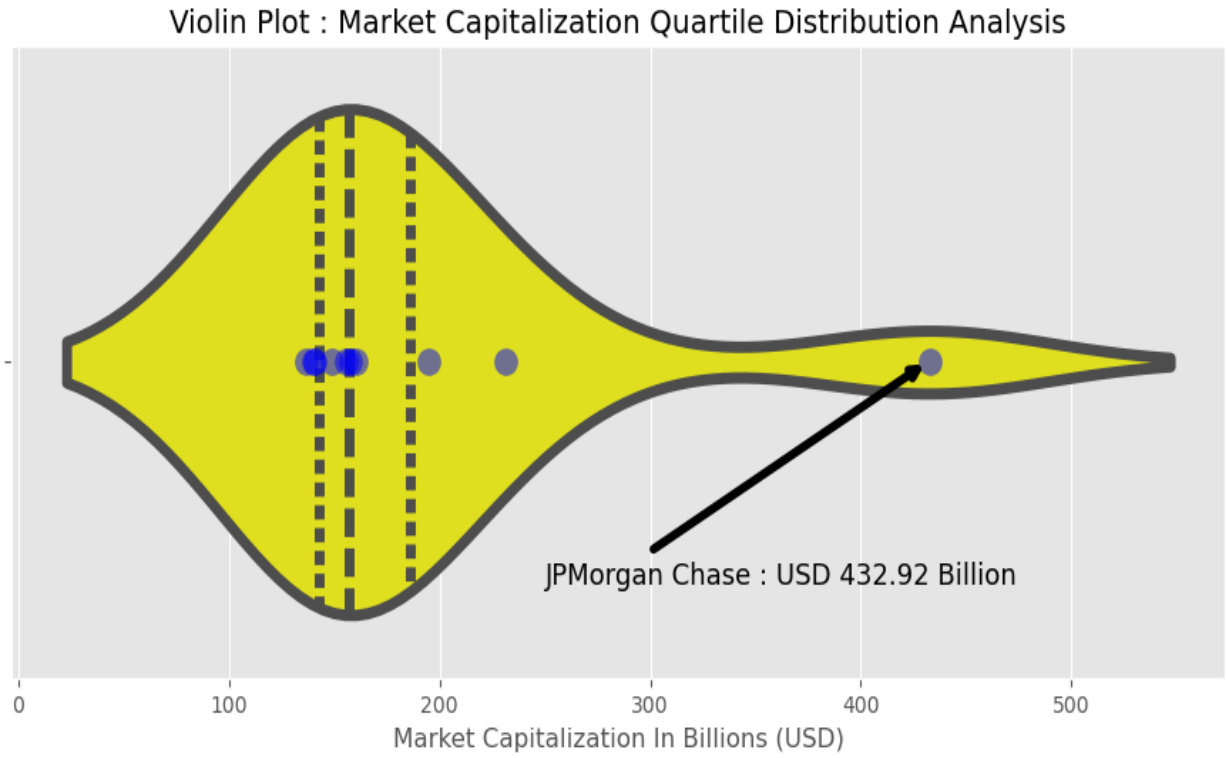
There is an extreme outlier to the extreme right of the chart as per the Box Plot (which leverages IQR method) but further analysis revealed that it was JPMorgan Chase, a market leader with a strong market capitalization.

The plot successfully validates the Outlier Analysis method performed during Data Cleaning stage was accurate.



3.6. Market Capitalization Quartile Distribution Analysis

A violin plot was used, since it is best for statistical representation of the spread and density of a numerical feature, in our case Market Capitalization. Each data point on the chart represents the market capitalization of an individual bank.



Quartile 1 (25%), Quartile 2 (Median) and Quartile 3 (75%) are represented on the plot as dashed lines. Majority of the data points are concentrated around Quartile 1 and Quartile 2 (Median), indicating lower to mid-capitalization range.

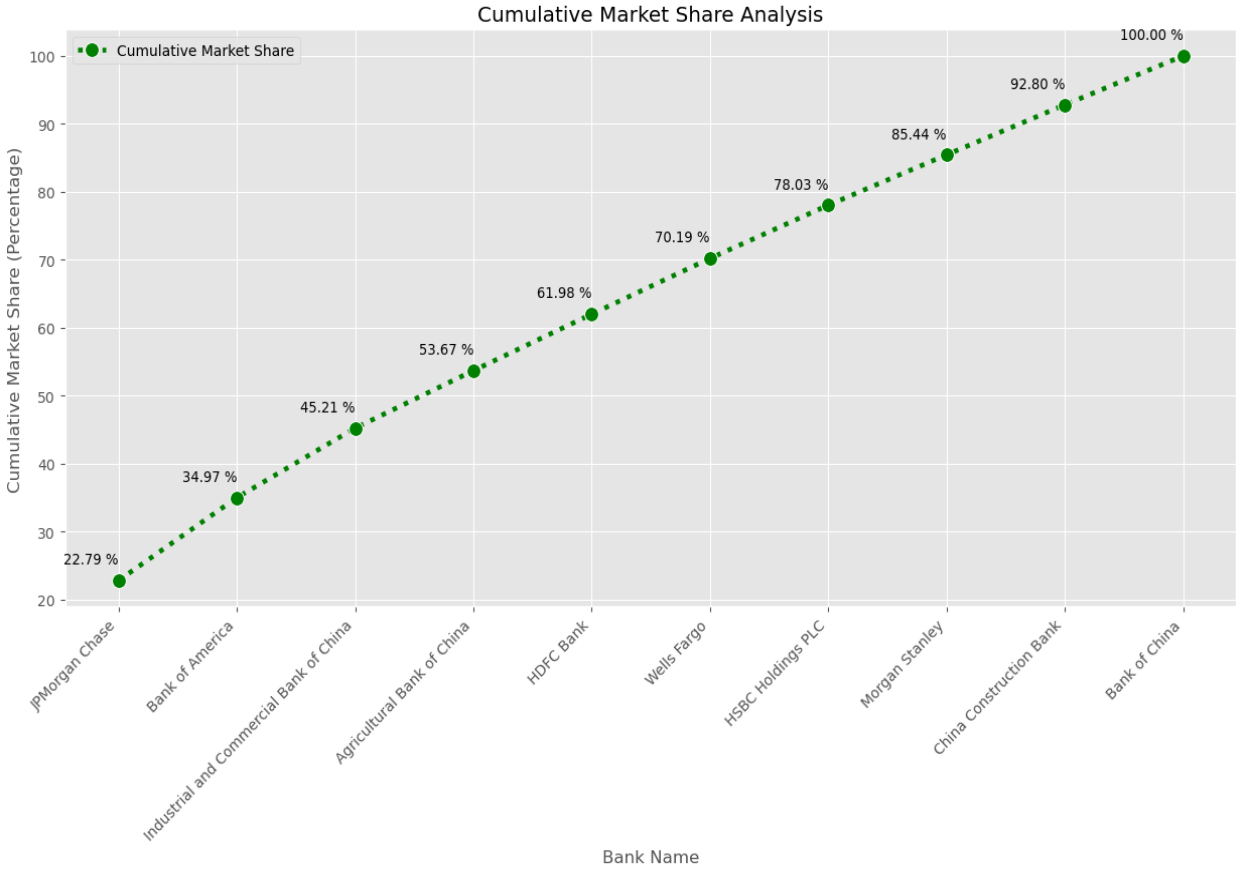
The distribution is skewed towards the right indicating that a few banking institutions have exceptionally high market capitalization (such as JPMorgan Chase). This also indicates the presence of dominant banks within the global banking industry.

Overall, the plot clearly indicates the presence of a few high-end institutions, whereas a vast majority of the players are mid-sized institutions, making it an effective and valuable tool for comparative market analysis.

3.7. Cumulative Market Share Analysis

Cumulative Market Share Analysis is useful for understanding the total market presence of companies as a group and how they dominate the market as a group rather than as an individual.

From the line plot generated it is evident that JPMorgan Chase holds the maximum market share as an individual institution (≈ 22.79%).



The Top 3 Banking Institutions have a control of over (≈ 45.21%) indicating a strong market dominance. Out of the Top 3 Banking Institutions JPMorgan Chase and Bank of America are from the United States.

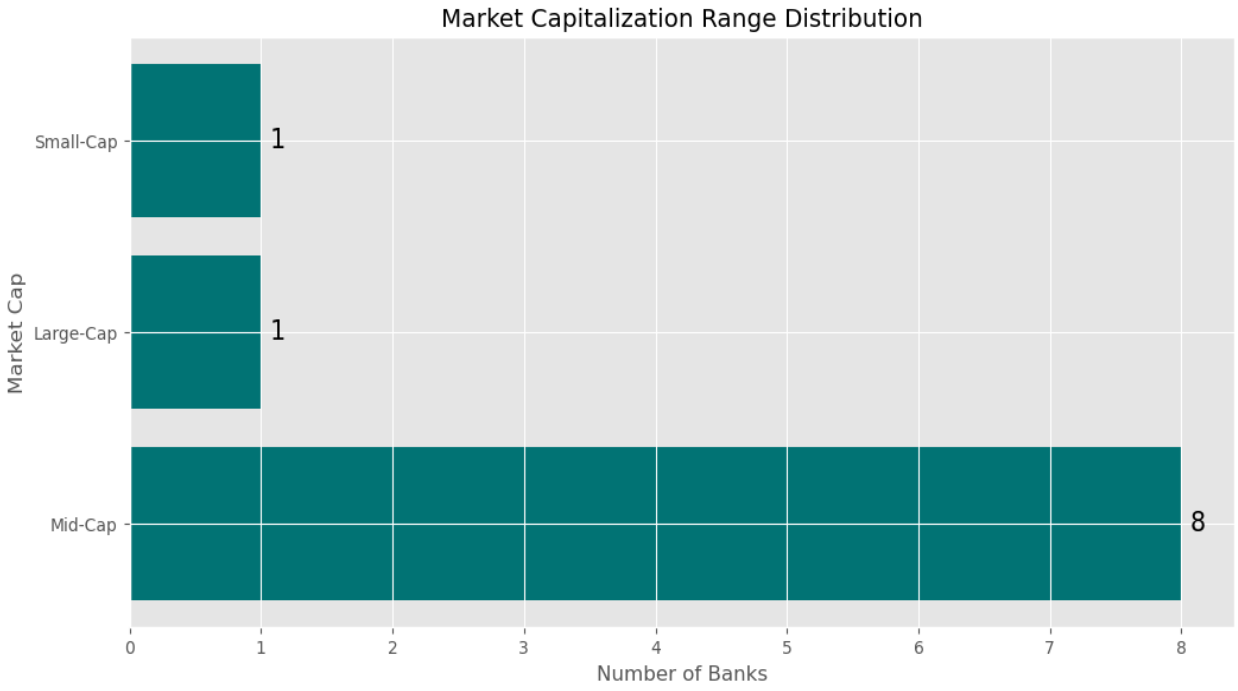
The chart clearly shows that just the Top 6 Banking institutions control over (≈ 70.19%), indicating a highly concentrated market. The plot clearly shows how a limited number of institutions command a large proportion of the global banking market capitalization.

3.8. Analyzing Market Capitalization Ranges and their Distribution Analysis using a bar chart.

A custom user-defined function was used to categorize banks into three main types, i.e. Small-Cap, Mid-Cap and Large-Cap based on their respective market capitalization. The logic written is scalable and doesn't depend on any fixed threshold for classification.

The objective is to understand different range segments and which segment dominates and which doesn't. As per our analysis majority of the banks are Mid-Cap Banks. With only a few Large-Cap and Small-Cap banks (1 each). This indicates that global banking institutions maintain a moderate market capitalization rather than being on either of the extreme ends of the spectrum.

Our findings indicate a high concentration of Mid-Cap Banks. Concluding that the broader global banking landscape is primarily composed of mid-sized players.

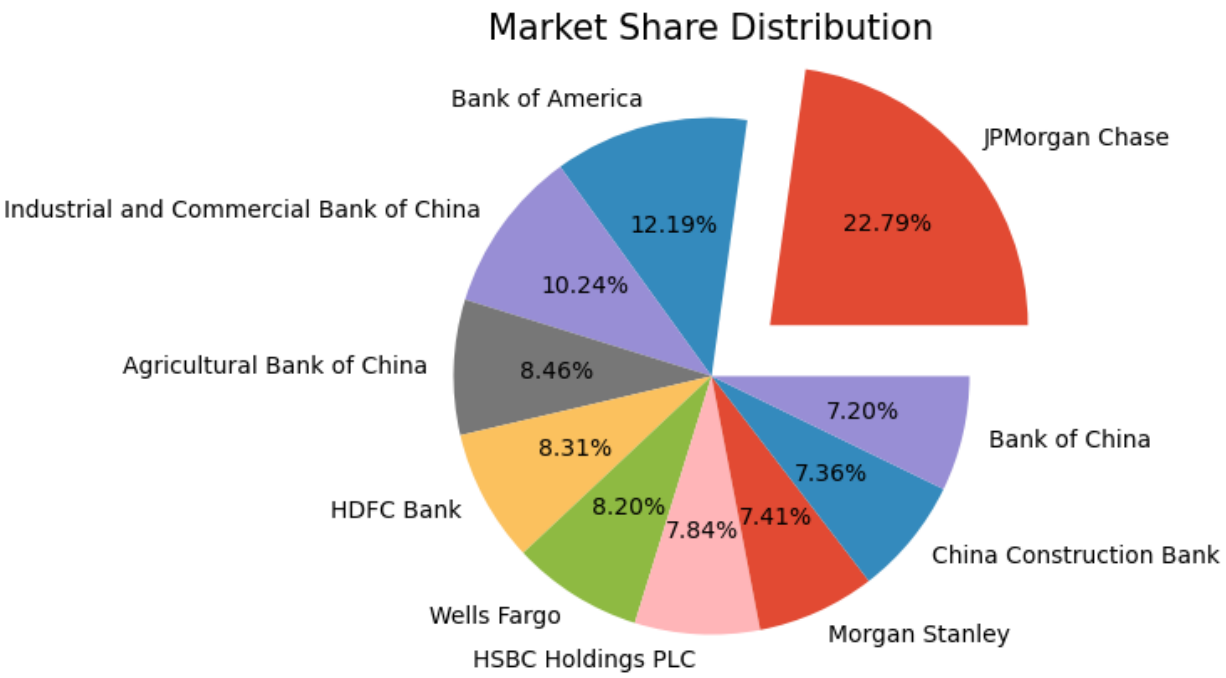


3.9. Market Share Distribution (Proportional Analysis)

It is important to understand that Pie-Chart are easily interpretated visually but it has some limitations; to overcome these limitations I have manually calculated and verified the percentages as well to ensure accuracy of the results. I have also displayed the respective percentage of the banks, making it easier to compare visually.

This distribution gives us a deeper dive into an individual market share of the global banking institutions.

The chart highlights dominance and uniformity among global banking institutions.



Majority of the banks hold a market share in the range of (≈ 7.20% to ≈ 8.46%) inclusive, reflecting a balanced representation among mid-tier institutions.

Only two Global Banking Institutions hold an individual market share in the range of (≈ 10.24% to ≈ 12.19%) inclusive.

It is also evident that JPMorgan Chase holds the largest market share as an individual institution (≈ 22.79%), positioning itself as market leader and dominant banking institution within our current dataset.

4. Banking Data ETL Querying

This section started with a critical transformation step of deriving a new feature dense rank. This new feature gives us a more consistent view of the performance as well as of the competition, making it an ideal feature for bank ranking. It provides us with a clear, consecutive and uninterrupted sequence which is essential for accurate analysis and reporting. This kind of feature plays a crucial role in analyzing competitiveness, dominance and comparative analysis.

4.1. Cross-Currency ETL Querying and Dashboarding

A specialized approach was used to ensure cross-currency ETL Querying and Dashboarding mechanism which plays an important role in performing cross-currency analysis.

In this approach the user gets to select the desired display currency out of the 4 currencies (USD, EUR, GBP, INR) before proceeding with further analysis or dashboard visualizations.

This step lays the foundation for identifying potential regions or banking segments for expansion and is useful for analyzing regional trends. It is also useful for banks as banks report their financial performance in different currencies depending on the location of their branch.

By embedding this logic in my ETL workflow, it ensures scalability and consistency and ensures meaningful international comparability. It also ensures that the dashboards stay relevant to the banks region thereby improving analytical consistency and integrity.

4.2. Advanced Market Capitalization Analysis with Growth Metrics.

This part of the ETL Querying is focused on advanced market capitalization analysis that incorporates various features (analytical indicators) for advanced analysis such as dense rank, individual market share, cumulative market share, gap from leading bank and growth rate.

Advanced Market Capitalization Analysis with Growth Metrics :-

DENSE_RANK	GLOBAL_RANK	BANK_NAME	MARKET_CAPITAL	MARKET_SHARE	CUMMULATIVE_MARKET_CAP	CUMMULATIVE_MARKET_SHARE	GAP_FROM_LEADING_BANK	GROWTH_RATE
1	1	JPMorgan Chase	432.92 Billion USD	22.79	432.92	22.79	0.0	127.87
2	2	Bank of America	231.52 Billion USD	12.19	664.44	34.98	-201.4	21.86
3	3	Industrial and Commercial Bank of China	194.56 Billion USD	10.24	859.0	45.22	-238.36	2.41
4	4	Agricultural Bank of China	160.68 Billion USD	8.46	1019.68	53.68	-272.24	-15.42
5	5	HDFC Bank	157.91 Billion USD	8.31	1177.59	61.99	-275.01	-16.88
6	6	Wells Fargo	155.87 Billion USD	8.2	1333.46	70.19	-277.05	-17.96
7	7	HSBC Holdings PLC	148.9 Billion USD	7.84	1482.36	78.03	-284.02	-21.62
8	8	Morgan Stanley	140.83 Billion USD	7.41	1623.19	85.44	-292.09	-25.87
9	9	China Construction Bank	139.82 Billion USD	7.36	1763.01	92.8	-293.1	-26.4
10	10	Bank of China	136.81 Billion USD	7.2	1899.82	100.0	-296.11	-27.99

The Growth Rate (Relative Performance Metric) was derived by comparing an individual bank's market capitalization against the banking industry average. This metric is essential for measuring how much a bank's market capitalization deviates from the standard industry average. This ensures consistency across all banks. Positive growth rate represents that banking institutions are performing above the industry average, whereas negative values highlight underperformance to the market benchmark.

JPMorgan Chase, Bank of America and Industrial and Commercial Bank of China are the only banks with a positive growth rate indicating a strong performance.

Gap from leading bank is a powerful analytical indicator which displays the difference between the current bank's market capitalization and the absolute market capitalization of the leader, expressed in the selected currency. It plays an important role in analyzing performance gaps between banks.

4.3. Market Concentration Analysis (On basis of Market Share Tiers)

Market Share Tiers is designed in a way to ensure scalability and consistency. It is purely quartile based and not based on any fixed threshold. This ensures dynamic categorization as the dataset increases in size, while maintaining analytical consistency.

Summary Table for Market Concentration Analysis and Threshold based on statistics:-

NUMBER OF BANKS	NAMES OF BANKS	MARKET_CONCENTRATION_TIERS	THRESHOLD
3	JPMorgan Chase, Bank of America, Industrial and Commercial Bank of China	MARKET LEADERS	> 9.8%
2	Agricultural Bank of China, HDFC Bank	STRONG PERFORMERS	8.25% - 9.8%
2	Wells Fargo, HSBC Holdings PLC	MID LEVEL PLAYERS	7.52% - 8.25%
3	Morgan Stanley, China Construction Bank, Bank of China	EMERGING PLAYERS	< 7.52%

Market Share Tiers classified banks into four main types on the basis of their individual market share. These four types are Market Leaders (3 Banks), Strong Performers (2 Banks), Mid-level Players (2 Banks) and Emerging Players (3 Banks).

Definition as per "Investopedia": The Herfindahl-Hirschman Index (HHI) measures the size of companies relative to the size of the industry they're in. It can range from close to zero to 10,000 with lower values, indicating a less concentrated market.

HHI Concentration Score to analyze how concentrated the market is:-

HHI_SCORE	HHI_SCORE_INTERPRETATION
1203.0	LOW CONCENTRATION - HIGHLY COMPETITIVE MARKET

As per our analysis, HHI Score was 1203, indicating a highly competitive market and low concentration. It is typically considered an unconcentrated market.

The combined use of Quartile based market share tier and HHI Score provides us with a comprehensive view on how much of the market is controlled by a small number of banking institutions. It can also be used to measure market concentration trends over time and as the size of the data increases.

It is important to note that this approach is adaptable across different currencies, allowing us to perform cross-currency evaluation. Banks can leverage the market concentration trends and can target growth opportunities.

4.4. Statistical Distribution of Market Capitalization Using Quartile Analysis.

Quartile based segmentation approach was leveraged while classifying global banking institutions into four quartile categories. This approach can dynamically adapt as the data changes in size, ensuring scalability.

The Quartile Analysis of Market Capitalization is as follows :-

QUARTILE_1 (Q1)	MEDIAN_QUARTILE_2 (Q2)	QUARTILE_3 (Q3)	INTERQUARTILE_RANGE (IQR)
142.85 Billion USD	156.89 Billion USD	186.09 Billion USD	43.24 Billion USD

bank_name	QUARTILE_CATEGORIES
JPMorgan Chase	FOURTH_QUARTILE
Bank of America	FOURTH_QUARTILE
Industrial and Commercial Bank of China	FOURTH_QUARTILE
Agricultural Bank of China	THIRD_QUARTILE
HDFC Bank	THIRD_QUARTILE
Wells Fargo	SECOND_QUARTILE
HSBC Holdings PLC	SECOND_QUARTILE
Morgan Stanley	FIRST_QUARTILE
China Construction Bank	FIRST_QUARTILE
Bank of China	FIRST_QUARTILE

Statistical Distribution Analysis is useful for analyzing the spread with the dataset and is also useful while looking for potential outliers. The analysis covers the Top 10 Banking institutions, with market capitalization ranging from USD 136.81 Billion to USD 432.92 Billion.

- **Quartile 1 (Q1) –** Bank of China, China Construction Bank and Morgan Stanley are emerging players with a market capitalization less than USD 142.85 Billion.
- **Quartile 2 (Q2) –** HSBC Holdings PLC and Wells Fargo are mid-level players with moderate market capitalization.
- **Quartile 3 (Q3) –** HDFC Bank and Agricultural Bank of China are strong performers who could be potential market leaders.
- **Quartile 4 (Q4) –** JPMorgan Chase, Bank of America and Industrial and Commercial Bank of China are market leaders and top performing institutions with a market capitalization above USD 186.09 Billion.
- **Interquartile Range (IQR) –** USD 43.24 Billion, indicating how spread 50% of the data is.

The Quartile Distribution Analysis also supports cross currency evaluation ensuring scalability for global comparisons. Quartile classification helps in tracking competitiveness among banks. Lower Quartile banks indicate potential areas for growth and expansion strategies can be implemented.

```
The Bank with the Least Market Capitalization is :-

+-------------+------------------+
|    BANK_NAME|MINIMUM_MARKET_CAP|
+-------------+------------------+
|Bank of China|136.81 Billion USD|
+-------------+------------------+

The Bank with the Highest Market Capitalization is :-

+--------------+------------------+
|     BANK_NAME|MAXIMUM_MARKET_CAP|
+--------------+------------------+
|JPMorgan Chase|432.92 Billion USD|
+--------------+------------------+
```

4.5. Comparative Size Analysis

The objective of this analysis was to classify banks on basis of their relative market size and to evaluate how large or small each banking institution is in relation to key market capitalization benchmarks.

The Four Market Capitalization benchmarks are comparison to leading bank, comparison to average, comparison to median and comparison to total market capitalization. Our focus will be on the first three benchmarks to avoid repeated analysis.

The Comparative Size Analysis: Compared to the Leading Bank Market Capitalization:-

global_rank	bank_name	RELATIVE_MARKET_SHARE	RELATIVE_SIZE
1	JPMorgan Chase	100.0	LARGE-SIZE
2	Bank of America	53.48	LARGE-SIZE
3	Industrial and Commercial Bank of China	44.94	LARGE-SIZE
4	Agricultural Bank of China	37.12	MEDIUM-SIZE
5	HDFC Bank	36.48	MEDIUM-SIZE
6	Wells Fargo	36.0	MEDIUM-SIZE
7	HSBC Holdings PLC	34.39	MEDIUM-SIZE
8	Morgan Stanley	32.53	SMALL-SIZE
9	China Construction Bank	32.3	SMALL-SIZE
10	Bank of China	31.6	SMALL-SIZE

The Comparative Size Analysis, which was compared to the leading bank market capitalization, measures the size of the bank relative to the largest banking institution. The Large-size global leaders, hold more than 40 percent of the leading benchmark value.

The Comparative Size Analysis: Compared to the Average Market Capitalization:-

global_rank	bank_name	RELATIVE_MARKET_SHARE	RELATIVE_SIZE
1	JPMorgan Chase	227.87	LARGE-SIZE
2	Bank of America	121.86	LARGE-SIZE
3	Industrial and Commercial Bank of China	102.41	LARGE-SIZE
4	Agricultural Bank of China	84.58	MEDIUM-SIZE
5	HDFC Bank	83.12	MEDIUM-SIZE
6	Wells Fargo	82.04	MEDIUM-SIZE
7	HSBC Holdings PLC	78.38	MEDIUM-SIZE
8	Morgan Stanley	74.13	SMALL-SIZE
9	China Construction Bank	73.6	SMALL-SIZE
10	Bank of China	72.01	SMALL-SIZE

The Comparative Size Analysis, which was compared to the average bank market capitalization, measures the size of each bank relative to the sector's overall mean size. The Top 3 Banks once again demonstrate significant advantage, with JPMorgan Chase's valuation more than twice the sector

average. Small-Size banks remain under 75% of the benchmark. The Mid-Size banking institutions suggest a tight competition.

The Comparative Size Analysis: Compared to the Median Market Capitalization:-

global_rank	bank_name	RELATIVE_MARKET_SHARE	RELATIVE_SIZE
1	JPMorgan Chase	275.94	LARGE-SIZE
2	Bank of America	147.57	LARGE-SIZE
3	Industrial and Commercial Bank of China	124.01	LARGE-SIZE
4	Agricultural Bank of China	102.42	MEDIUM-SIZE
5	HDFC Bank	100.65	MEDIUM-SIZE
6	Wells Fargo	99.35	MEDIUM-SIZE
7	HSBC Holdings PLC	94.91	MEDIUM-SIZE
8	Morgan Stanley	89.76	SMALL-SIZE
9	China Construction Bank	89.12	SMALL-SIZE
10	Bank of China	87.2	SMALL-SIZE

The Comparative Size Analysis, which was compared to the median of bank's market capitalization, measures the size of each bank relative to the market's midpoint. It reveals positive skewness in size distribution of the global banking market. Meaning a small number of institutions hold a significantly large proportion, emphasizing dominance at the top-end.

The Market Comparative Size Analysis system remains scalable and supports cross-currency evaluation and doesn't classify sizes using fixed threshold but instead leverages a dynamic 3-tier percentile classification approach.

4.6. Market Growth and Gap Analysis

This kind of analysis is important to evaluate the growth performance and competitive gaps between global banking institutions on the basis of their market capitalization. This analysis is scalable and supports cross-currency evaluation of market capitalization.

GLOBAL_RANK	BANK_NAME	MARKET_CAPITAL	GAP_TO_NEXT	GAP_PERCENTAGE	GROWTH_RATE	OUTPERFORMANCE_RATE	GAP_VELOCITY	MARKET_EXPANSION_INDEX
1	JPMorgan Chase	432.92 Billion USD	201.4	46.52	127.87	35.89	100.0	92.78
2	Bank of America	231.52 Billion USD	36.96	15.96	21.86	5.33	18.35	15.81
3	Industrial and Commercial Bank of China	194.56 Billion USD	33.88	17.41	2.41	6.78	16.82	6.3
4	Agricultural Bank of China	160.68 Billion USD	2.77	1.72	-15.42	-8.91	1.38	-10.42
5	HDFC Bank	157.91 Billion USD	2.04	1.29	-16.88	-9.34	1.01	-11.35
6	Wells Fargo	155.87 Billion USD	6.97	4.47	-17.96	-6.16	3.46	-10.42
7	HSBC Holdings PLC	148.9 Billion USD	8.07	5.42	-21.62	-5.21	4.01	-11.85
8	Morgan Stanley	140.83 Billion USD	1.01	0.72	-25.87	-9.91	0.5	-16.12
9	China Construction Bank	139.82 Billion USD	3.01	2.15	-26.4	-8.48	1.49	-15.75

Gap calculations were performed based on the descending sorted order of market capitalization of banks.

The Key Performance Indicators used in this analysis are as follows :-

- **Gap to Next –** This indicator helps in measuring the absolute difference in market capitalization between consecutive banking institutions.

- **Gap Percentage –** This indicator represents gaps between current and consecutive banking institutions in terms of percentage normalized by the current market capitalization.

- **Growth Rate –** This is a comparative performance indicator essential for measuring how much a bank's market capitalization deviates from the standard industry average.

- **Outperformance Rate –** This indicator is used to determine how much a bank's dominance exceeds or falls below the market average. Useful in determining the presence of banking institutions.

- **Market Expansion Index –** This custom-weighted indicator combines growth rate, outperformance rate and gap velocity features to represent the bank's overall market expansion capability and growth potential. Serving as a primary indicator for strategic planning.

JPMorgan Chase has the highest Market Expansion Index (MEI) and a Gap Velocity of 100%. This indicates a strong presence, massive competitive lead and strong growth momentum. No other bank comes close to this level.

Morgan Stanley and China Construction Bank have the lowest MEI and gap velocities. This indicates a weaker market presence and less chances for expansion.

Bank of America and Industrial & Commercial Bank of China maintain positive growth rates, outperformance rate and market expansion index. This indicates a strong growth position and strong presence among other banking institutions.

From Agricultural Bank of China onward, the momentum is negative. This indicates limited potential for expansion and below-average performance. Many of them maintain mid-range positions with lesser gaps indicating tight competition and still have some level of market influence.

This kind of analysis is useful for cross-currency evaluation, monitoring concentration and growth trends and identifying potential expansion opportunities.

4.7. Market Dominance Analysis (Market Dominance Score)

This section of the analysis focuses on analyzing the market dominance of banks among other global banking institutions on the basis of their individual and cumulative market dominance score.

The Two main market dominance indicators are as follows:-

- **Individual Dominance Score (Micro Level)–** This indicator is used for measuring the dominance level of each bank as an individual. It measures how far each bank's share deviates from the fair share. Stronger score represents greater power compared to the other banking institutions.
- **Cumulative Dominance Score (Macro Level) –** This indicator is used for measuring the dominance level of banks as a group. It shows how dominance accumulates as we move down the global banks based on their global ranks. Higher cumulative dominance score represents a greater concentration of market share at the top level of banks. It compares the cumulative market share with the expected cumulative market share.

Both the indicators are designed to operate on normalized scales.

Cummulative Dominance Score and Individual Dominance Score and Cummulative Market Share are as follows:-

global_rank	bank_name	MARKET_SHARE	CUMMULATIVE_MARKET_SHARE	INDIVIDUAL_DOMINANCE_SCORE	CUMMULATIVE_DOMINANCE_SCORE
1	JPMorgan Chase	22.79	22.79	5.71	2.28
2	Bank of America	12.19	34.98	5.12	2.69
3	Industrial and Commercial Bank of China	10.24	45.22	5.01	2.96
4	Agricultural Bank of China	8.46	53.68	4.91	3.05
5	HDFC Bank	8.31	61.99	4.91	3.16
6	Wells Fargo	8.2	70.19	4.9	3.29
7	HSBC Holdings PLC	7.84	78.03	4.88	3.41
8	Morgan Stanley	7.41	85.44	4.86	3.45
9	China Construction Bank	7.36	92.8	4.85	3.52
10	Bank of China	7.2	100.0	4.84	NULL

JPMorgan Chase has the highest individual and cumulative dominance score which indicates the institution as a market leader with a strong influence over the global banking market and a strong dominance.

Bank of America and Industrial & Commercial Bank of China, both maintain a strong dominance score individually and as a group. When these two banking institutions are combined with JPMorgan Chase, they control over 45.22% of the total market share.

The decline in the cumulative dominance score after the top 3 banking institutions indicate that all the institutions are equal after that point. Rank 3 can be considered as the point of maximum market concentration.

4.8. Segment-Wise Bank Performance Analysis

This section focuses on analyzing segment-wise bank performance by classifying them into three segments : Large-Cap, Mid-Cap and Small-Cap.

The classification is not based on fixed threshold but is scalable as it uses *np.linspace* method which ensures adaptability and scalability as the size of the data grows ensuring analytical consistency and cross-currency evaluation.

- **Large-Cap –** The top-performing banks with the highest market capitalization.
- **Mid-Cap –** The institutions with a moderate to high market capitalization. Indicating stable market performance and tight competition.
- **Small-Cap –** The institutions with a relatively small market capitalization. Indicating potential growth opportunities.

Large-Cap Segment Analysis : JPMorgan Chase is the only institution which is classified as Large-Cap, it contributes to a significant market share of 22.79%. It positions itself as a global market leader with a strong influence over the market.

Mid-Cap Segment Analysis : Majority of the banks fall in this segment. This is a strong indication of tight competition. Mid-Cap segment contributes to 70.01 % of the total market capitalization.

Small-Cap Segment Analysis : Bank of China falls in this segment contributing 7.2 % of the total market share. This is an indication for potential growth opportunities if implemented strategically.

Let's view the Top 10 banks and their Market Cap Tier (in USD):

bank_name	MARKET_CAPITALIZATION_BILLION	market_cap_tiers
JPMorgan Chase	432.92	Large-Cap
Bank of America	231.52	Mid-Cap
Industrial and Commercial Bank of China	194.56	Mid-Cap
Agricultural Bank of China	160.68	Mid-Cap
HDFC Bank	157.91	Mid-Cap
Wells Fargo	155.87	Mid-Cap
HSBC Holdings PLC	148.9	Mid-Cap
Morgan Stanley	140.83	Mid-Cap
China Construction Bank	139.82	Mid-Cap
Bank of China	136.81	Small-Cap

Let's view the Segment-Wise Performance Detailed Analysis :

MARKET_CAP_TIERS	NUMBER_OF_BANKS	AVERAGE_MARKET_CAP	MINIMUM_MARKET_CAP	MAXIMUM_MARKET_CAP	MARKET_SHARE_PERCENTAGE
Large-Cap	1	432.92 Billion USD	432.92 Billion USD	432.92 Billion USD	22.79 %
Mid-Cap	8	166.26 Billion USD	139.82 Billion USD	231.52 Billion USD	70.01 %
Small-Cap	1	136.81 Billion USD	136.81 Billion USD	136.81 Billion USD	7.2 %

4.9. Comprehensive Performance Dashboard for Bank Rankings and Metrics.

The dashboard supports cross-currency evaluation and ensures scalability and easy access to market performance, concentration, market capitalization distribution and growth rate among the few analytical indicators.

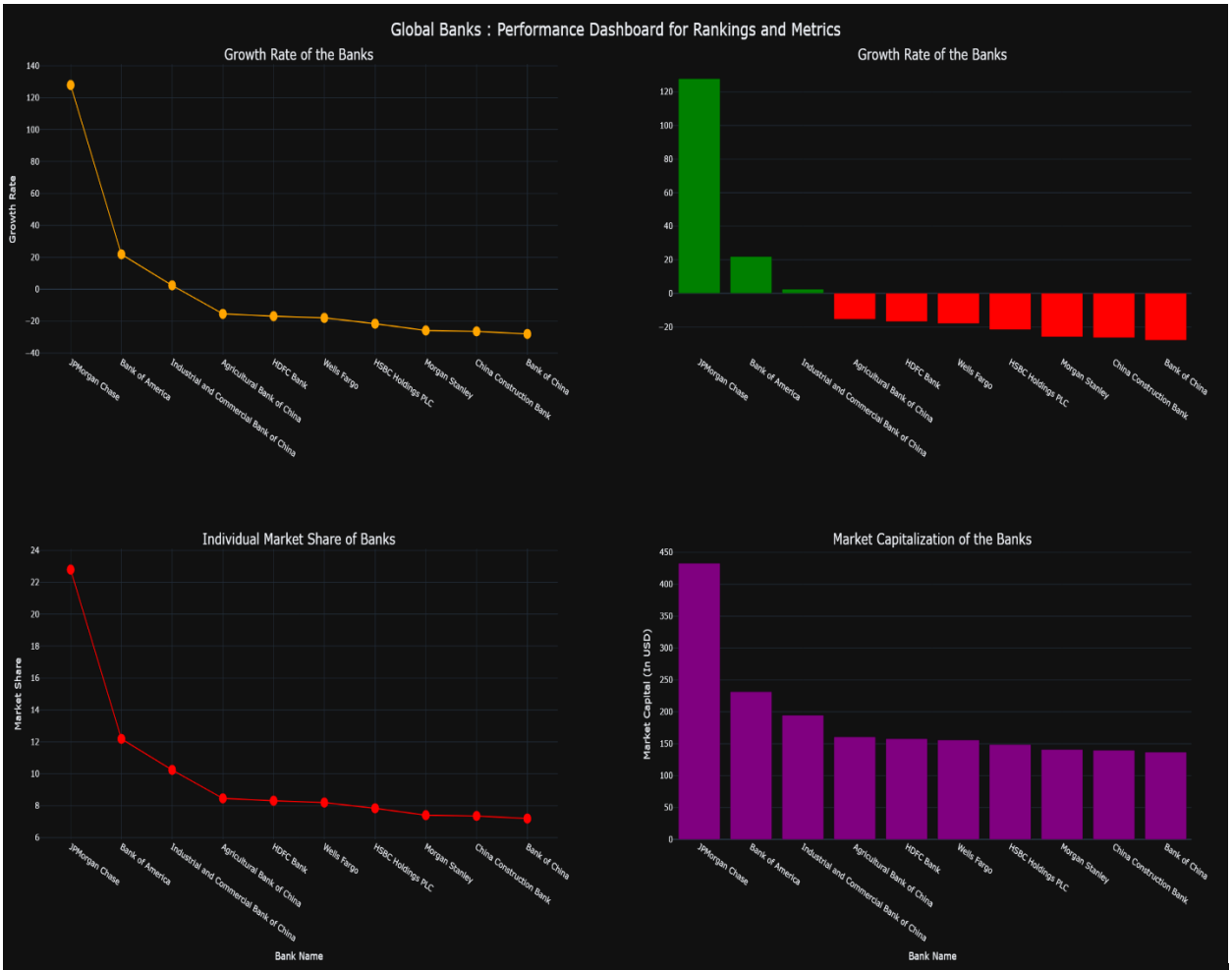
The dashboard provides a view into the market's current condition. It serves as the foundation for regional managers to evaluate the market at a glance enabling effective and efficient market evaluation and strategic planning. The KPI's provide a summary of the market structure and performance.

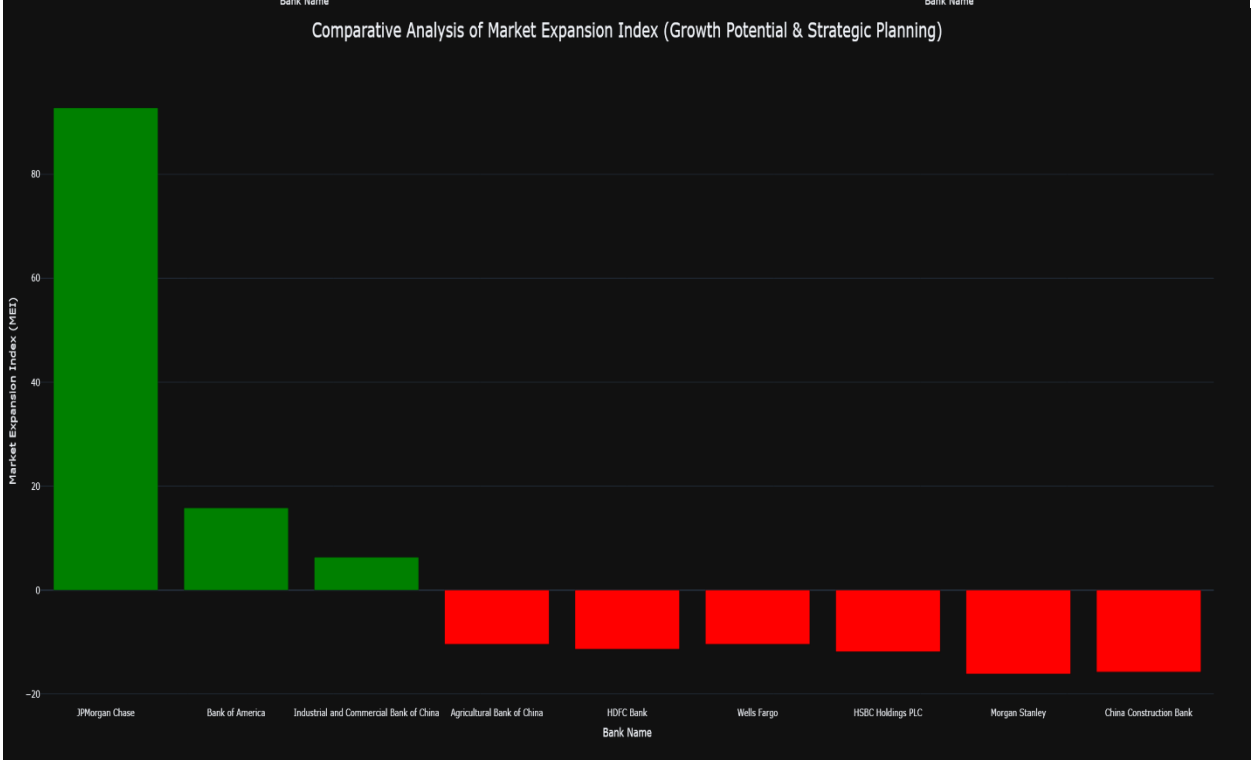
The Key Performance Indicators include some essential key metrics and address critical analytical questions such as:

- What is the Total Market Capital?
- How many global banks are there in the market?
- What is the level of concentration on the basis of HHI Score?
- Who is the market leader and what is their growth rate and how much market capitalization do they have?
- Which bank has potential to grow, and currently has the least market capitalization?

The performance dashboard provides a deeper analytical visual of growth rate, market share, market capitalization and market expansion index. Red Bars in the dashboard indicate negativity.

- **Growth Rate :** Illustrates growth trend among different banking institutions, showing how performance varies.
- **Market Capitalization Distribution :** Displayed via a bar chart, showing how a group of banks (Top 3 Banks) dominate the overall market capitalization.
- **Individual Market Share :** Displays how the market share is distributed among different banking institutions (individually). There is a visible drop after the top three banks indicating dominance & competitive imbalance.
- **Market Expansion Index (MEI) :** Represents the bank's overall market expansion capability and growth potential. Thereby, serving as a primary indicator for strategic planning and positioning.


Global Banks : Performance Dashboard for Rankings and Metrics
Growth Rate of the Banks
Growth Rate
Growth Rate of the Banks
Individual Market Share of Banks
Market Share
Market Capitalization of the Banks
Market Capital (in USD)
JPMorgan Chase
Bank of America
Industrial and Commercial Bank of China
Agricultural Bank of China
HDFC Bank
Wells Fargo
HSBC Holdings PLC
Morgan Stanley
China Construction Bank
Bank of China
Bank Name


Comparative Analysis of Market Expansion Index (Growth Potential & Strategic Planning)
Market Expansion Index (MEI)
JPMorgan Chase
Bank of America
Industrial and Commercial Bank of China
Agricultural Bank of China
HDFC Bank
Wells Fargo
HSBC Holdings PLC
Morgan Stanley
China Construction Bank
Bank Name

5. Conclusions (Final Insights and Recommendations)

Collectively, these data-driven recommendations derived from advanced analytics such as advanced growth rate, statistical distribution, market dominance, HHI, MEI and other indications provide a scalable cross-currency and analytical consistent framework for banking institutions across the globe. This unified model will help key decision makers make informed decisions based on real-time analytics.

5.1. Recommendations to track and compare market capitalisation of the top global banks to evaluate competitiveness and dominance.

- Market Capitalization is the primary and most fundamental metric for evaluating financial strength, competitiveness, concentration and dominance of global banking institutions. The dashboard and other plots are a definitive proof of why it is the most direct indicator of market power.

- The Market Expansion Index (MEI) and Growth Rate metrics should be integrated into the monitoring framework; it plays a crucial role in tracking the growth among different global banks individually as well as a whole. MEI indicator is important for strategic planning and analyzing the potential of expansions.

- The Market Dominance Score a powerful metric for tracking the dominance of banks. Both Individual Dominance Score (IDS) and Cumulative Dominance Score (CDS) were derived using market capitalization.

- For analytical consistency, I highly recommend that quarterly market capitalization tracking must be implemented so that the data can be monitored on a timely basis. This would ensure quarterly updates and reflect changes over time.

- This recommendation is data driven based on our key findings from our Segment-wise Analysis, Statistical Distribution Analysis, Market Dominance Analysis and our Visually Integrated Dashboard.

5.2. Suggestions to use cross-currency analysis (USD, GBP, EUR, INR) for consistent benchmarking of financial institutions across regions.

- The ETL pipeline was built to solve cross-currency evaluation problem. It was designed to ingest exchange rate data of different currencies and translate the single base currency (USD) into relevant user-selected currency for analysis (USD, GBP, EUR, INR). This is the foundational step before performing any ETL Querying or Dashboarding.

- Cross-currency analysis plays a crucial role in evaluating banking institutions across the globe. Since banks operate in different regions and currencies, relying on a single-currency view (such as USD) may distort the true comparative insights.

- This kind of analysis is important for global banking institutions to hedge foreign exchange risk, manage funding costs, handle operational inefficiencies in international trade & investments and high borrowing costs in foreign market.

- The cross-currency view offers a broader insight into global banking market capitalization. It enables traders, investors, managers and a vast group of people to analyze and capitalize on the economic changes between specific regions or countries with relevant information of different currencies and their volatility. This in turn helps in making more informed decisions in business, investment and trading.

- The ETL model should be leveraged in achieving exchange rate normalization across different currencies and evaluated on a daily, monthly or quarterly basis to ensure analytical accuracy and consistency and minimize exchange rate volatility during comparative analytics. This kind of model provides a unified analytical framework, to ensure consistency with global benchmarking standards such as the International Financial Reporting Standards (IFRS).

5.3. Propose continuous monitoring of market share concentration to identify growth opportunities for mid-tier banks.

- Continuous monitoring of market share concentration using this ETL pipeline is essential for evaluation of the global banking industry in terms of market concentration, competitiveness, dominance, growth and how different segments perform.

- Market Concentration is a key metric for observing how the market is controlled and distributed among the leading and emerging global banking institutions.

- A study on the Market Concentration using Herfindahl-Hirschman Index (HHI) revealed an HHI score of 1203, indicating a highly competitive marketplace and very little concentration. Despite the influence of the top three banking institutions (JPMorgan Chase, Bank of America and Industrial & Commercial Bank of China) which collectively hold over 45% of the total market share.

- I highly recommend that market share concentration should be measured on a timely basis in a continuous manner using both the HHI metric and Quartile-based segmentation. Tracking these metrics over time can help in detecting early signs of market concentration, dominance or weakening of competition or potential crashes.

- Our HHI Score of 1203, indicates a highly competitive and unconcentrated market, is powerfully reinforced in our Segment-Wise Analysis which revealed that 80% of the bank (8 out of 10 banks) controlled over (≈ 70.01%) of the total market share and are classified as Mid-Cap banks. This proves that Mid-Cap field is the most competitive, primary and most crowded in terms of market share. This indicates potential growth opportunities.

- Since it is a highly competitive market based on the HHI score, it encourages innovation, stability and efficiency across global banking institutions. These insights are crucial as they can help aid financial regulators, policy makers and key decision maker make informed decision about early signal of market imbalance or systemic risks or exchange rate volatility.

5.4. Identify potential regions or banking segments for expansion by analyzing gaps between tiers of banks and regional trends.

- The main indicator used to analyze the expansion is Market Expansion Index (MEI) combined with Gap Analysis. These indicators play a crucial role in identification of banks in different tiers and their respective regional performance trends. The banks with a negative MEI score can be targeted for strategic expansion planning.

- Our analysis indicates that Large-Cap Banking institutions have achieved strong market dominance and hold a significant share of the global market capitalization.

- Mid-Cap institutions have shown stable but moderate performance compared to Large-Cap Banking institutions and position themselves as ideal and competitive candidates for strategic regional expansion as well as diversification.

- Small-Cap banking institutions have shown limited market presence but by analyzing the right metrics and following a strategic and structured approach with the aim for targeted growth and consistent growth over time they have the potential to evolve.

- I highly recommend conducting regional performance tracking using Market Expansion Index and growth gap analytics. The focus should be towards tapping the untapped potential such as emerging Asian and Middle Eastern markets. Mid-tier banks have the potential to bridge the gap between top performer and establish a strong presence in the global market.

- The cross-currency analysis is the critical analytical tool for the recommendation of expansion. It allows key decision makers to leverage information based on region and local currency benchmarks to uncover insights related to performance, expansion and growth which would be invisible in a USD-only view. Regional diversification is important as it reduces dependency on a single economic zone or currency and helps mitigate exposure and risk.

Note: *The report primarily references USD as the base currency for all analytics. While the ETL pipeline is designed to support cross-currency evaluation (USD, GBP, EUR, INR). USD was used to maintain academic integrity and ensure consistency and comparability.*