1. **Data Preparation**

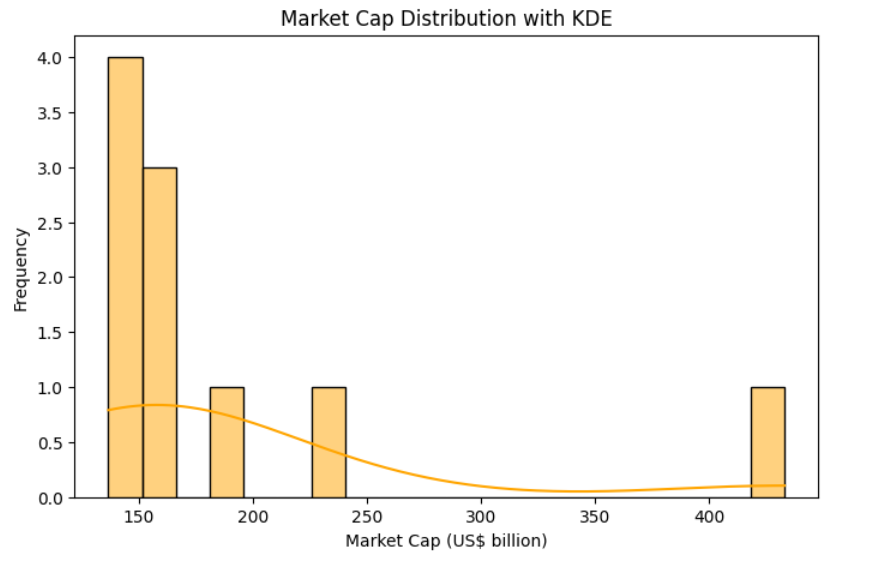
The dataset is scraped from the Wikipedia page provided and the table has been extracted.

1. **Data Cleaning**

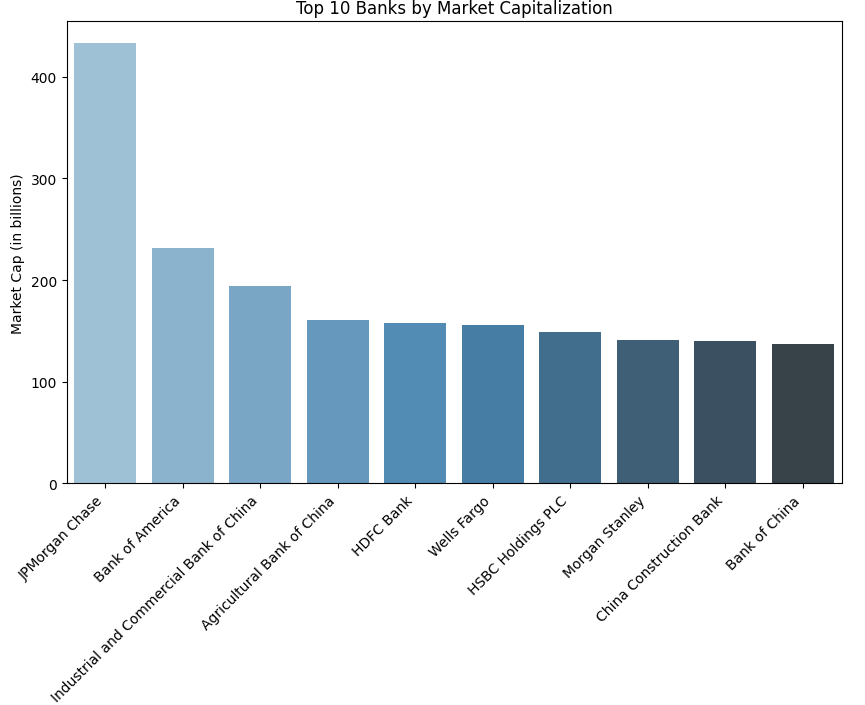
2.1. **Handling Missing Values:** There are no missing values  
2.2. **Fixing Columns:** All columns are properly named and formatted.  
2.3. **Handling Outliers:** Using quartile approach, we see JP Morgan Chase is an outlier indicating that it’s a clear leader based on it’s market share.

1. **Exploratory Data Analysis:**

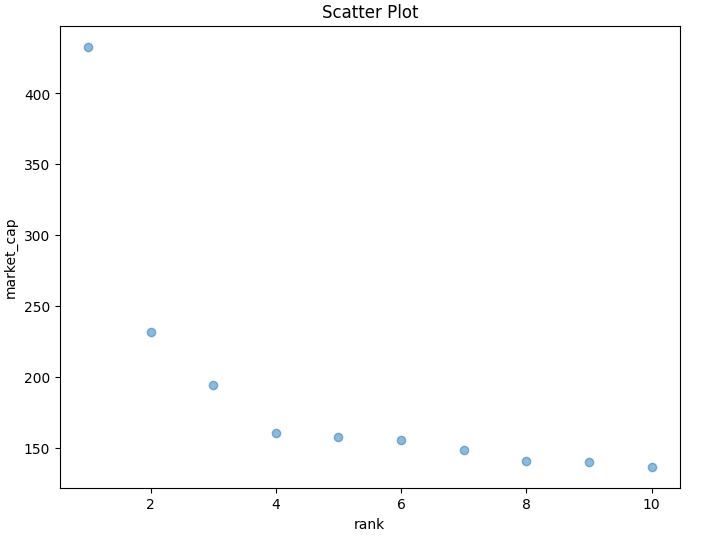
3.1. PySpark DataFrame is converted to Pandas DataFrame for visualization.  
3.2. Using a histogram to analyze the distribution of market capitalization



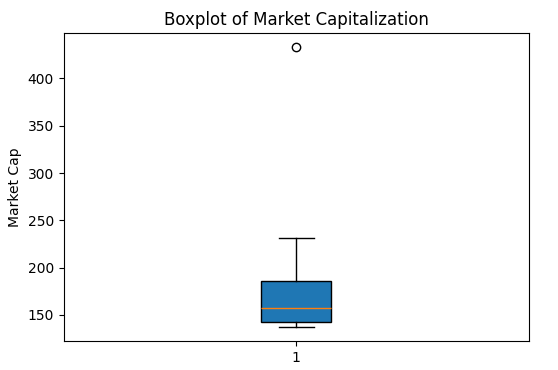
3.3. Identify the top 10 banks by market capitalization using a bar chart.

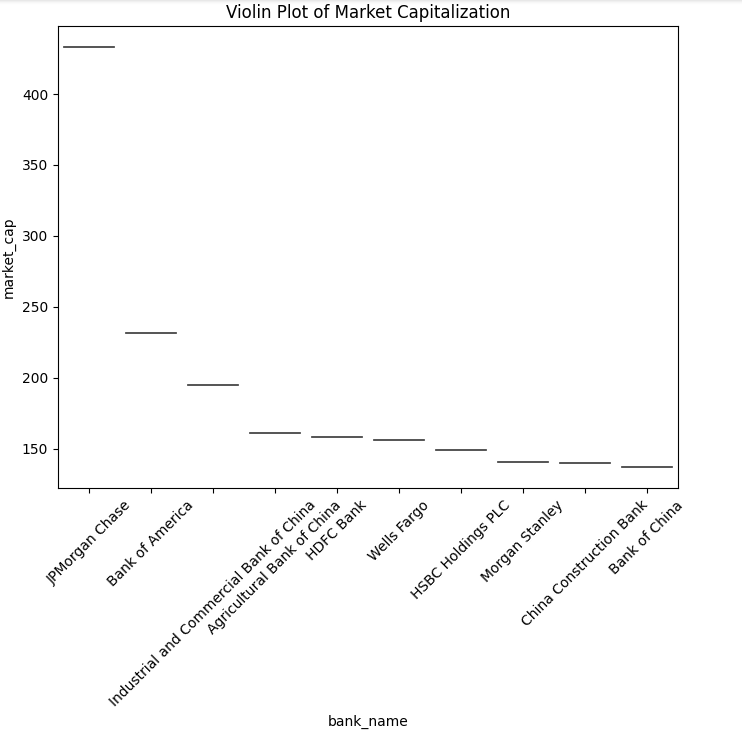


3.4. Using a scatter plot to visualize the relationship between market capitalization and bank ranking

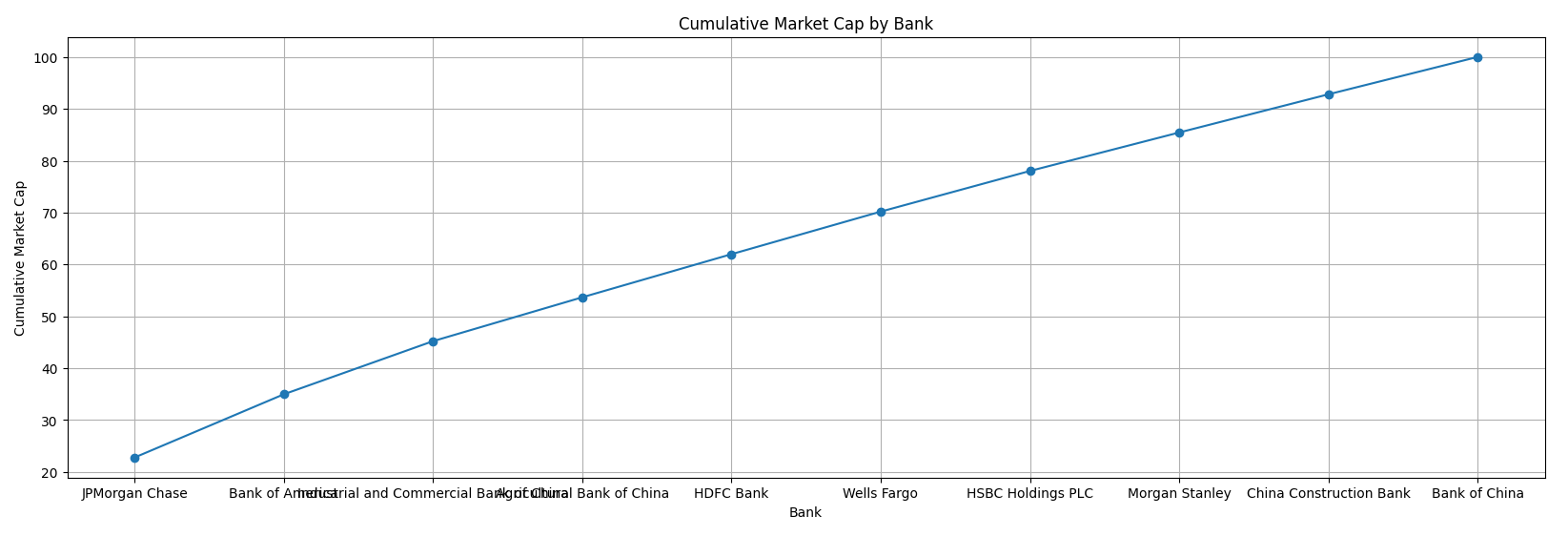


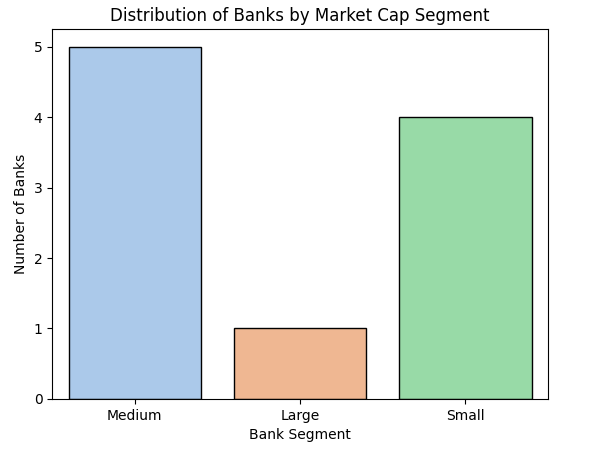
3.5. Boxplot to examine the spread and outliers in market capitalization.

  
3.6. Display the quartile distribution of market capitalization using a violin plot.

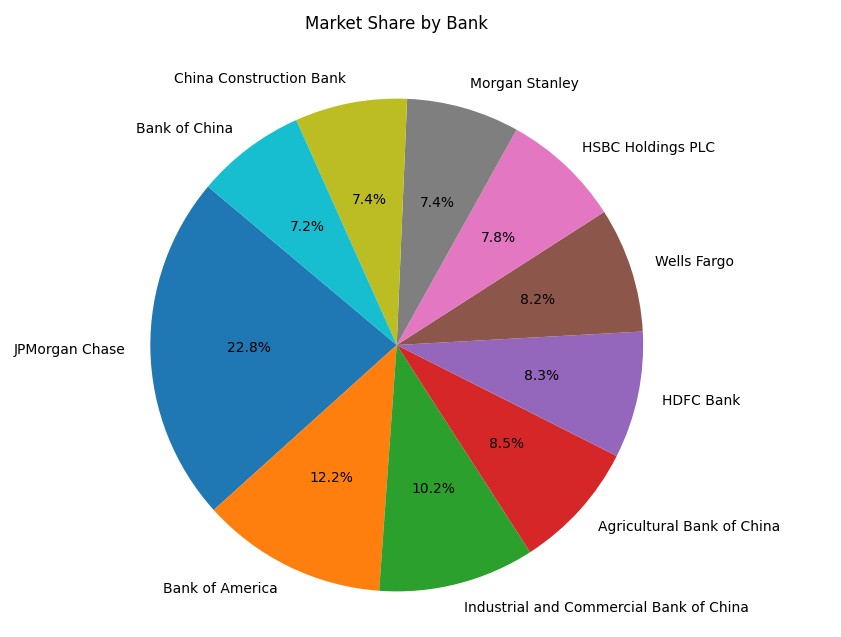


3.7. Compute cumulative market share and visualize it with a line plot.

  
3.8. Categorize banks into market capitalization ranges and analyze their distribution using a bar chart.



3.9. Calculate and display market share distribution of top 10 banks using a pie chart.

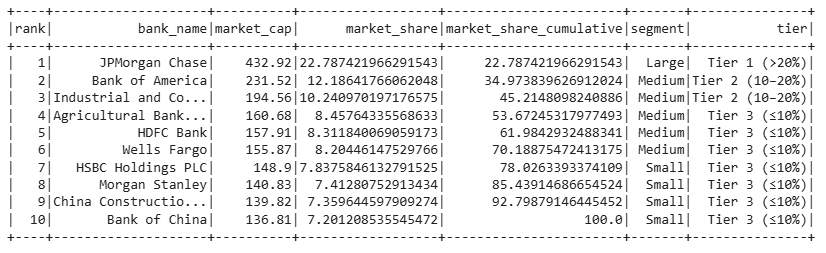


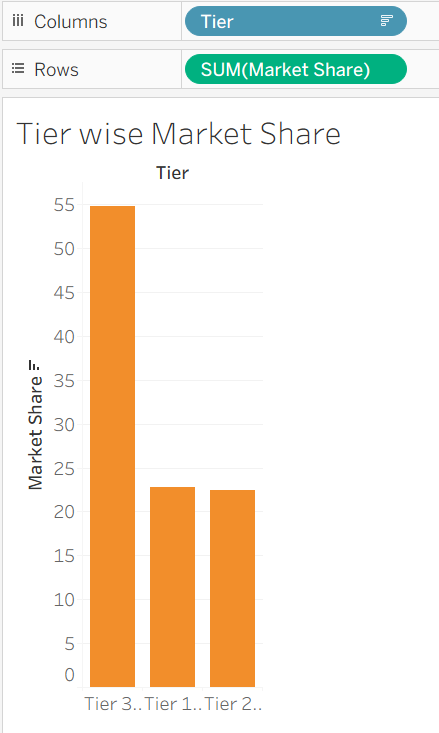
1. **Banking Data ETL Querying**

4.1. Perform Advanced Market Capitalization Analysis with Growth Metrics.

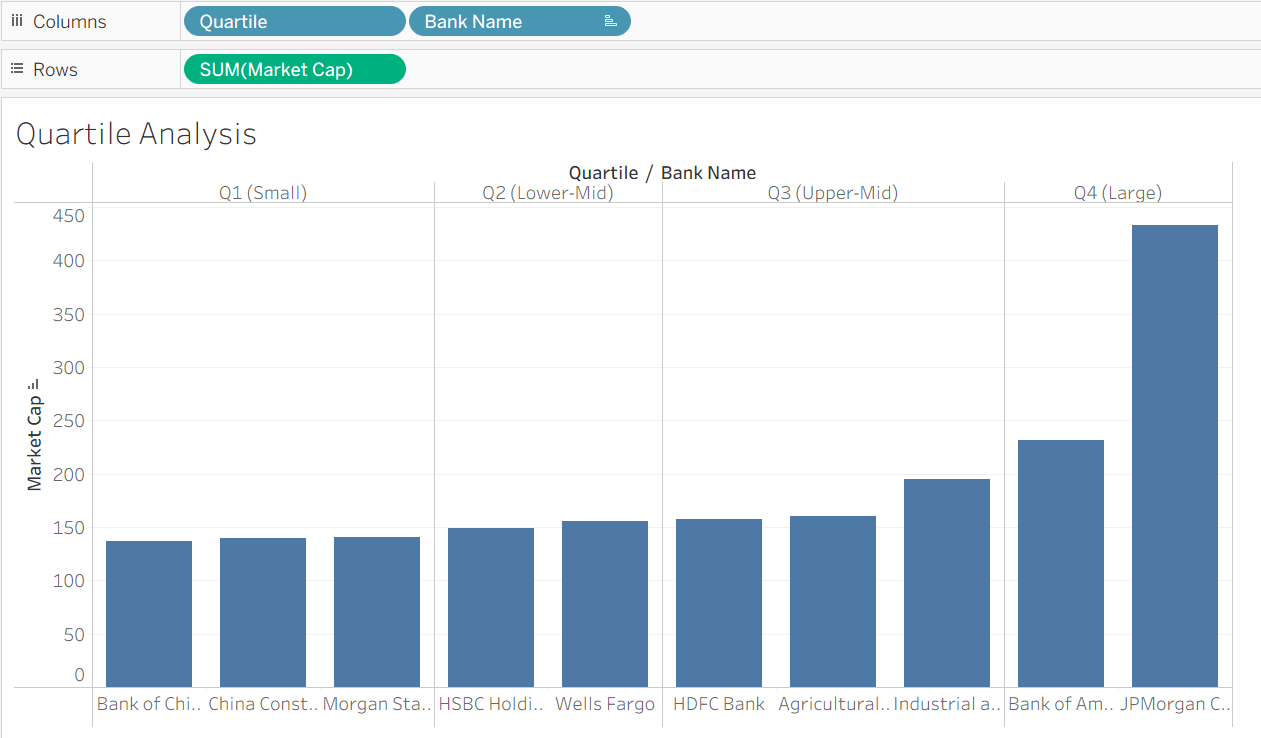
As growth metrics requires timeseries data, the table information is insufficient for analysis.

4.2. Analyze Market Concentration and Categorize Banks Based on Market Share Tiers.





4.3. Examine Statistical Distribution of Market Capitalization Using Quartile Analysis.



4.4. Conduct Comparative Size Analysis to Classify Banks by Relative Market Size.

The market share % can be used to classify banks as small, medium, large.

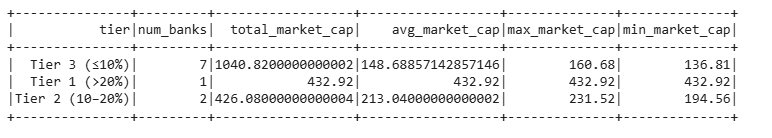
4.5. Evaluate Market Growth and Identify Gaps Between Consecutive Banks.

4.6. Assess Market Dominance by Measuring Cumulative Share and Dominance Score.

Herfindahl–Hirschman Index (HHI) : 0.120 (Normalized value)

This index is a measure used in **economics and finance** to understand how concentrated or competitive a market is. Calculated by the sum of squares of the market share of all banks.

4.7. Analyze Segment-Wise Bank Performance Based on Market Capitalization Ranges.



4.8. Generate a Comprehensive Performance Dashboard for Bank Rankings and Metrics.

