Documentation of ETL pipeline:

The pipeline begins by importing necessary libraries, including pandas for data manipulation, matplotlib and seaborn for visualization, and sqlite3 for database operations. It then installs the pandas library to ensure it's available for use. The two CSV files that are used as data sources: 'Cost_of_Living_Index_by_Country_2024.csv' and 'world-data-2023.csv'. These files are read into pandas DataFrames using ‘pd.read_csv()’. The transformation step involves merging the two DataFrames based on the ‘Country’ column. This combines the cost of living data with world demographic and economic data for countries present in both datasets. The ‘inner’ join ensures that only countries present in both datasets are included in the final result. The transformed data is then saved to a new CSV file and this step creates a new CSV file named ‘merged_dataset.csv’ in the ‘DATA’ directory, containing the combined and transformed data. Finally, the ‘merged_dataset.csv’ is stored into the MongoDB database under the name of ‘DS2002FinalProject’ and the collection of ‘Cost_of_living’.

Documentation of Google Cloud:

The setup involves creating a Google Cloud Storage bucket named ds2002-final-project, configured with a multi-region location in the United States and a default storage class of Standard. This configuration ensures high availability and durability of the stored data, making it accessible across multiple data centers within the region. By choosing the Standard storage class, the setup is optimized for frequently accessed data, which aligns with the project's need to retrieve transformed data for analysis efficiently. Access control is set to uniform, simplifying the permission management process by applying consistent policies across all objects in the bucket. The encryption is managed by Google, ensuring that data is securely stored without the need for manual encryption key management. This setup facilitates the reliable and secure storage of transformed data, making it readily available for further analysis and processing steps in the project.