During our ETL setup and implementation, we encountered a challenge when integrating two datasets with differing scopes. Our primary dataset, world-data-2023.csv, contains information on 195 countries, while the supplementary dataset, Cost of Living Index by country.csv, covers only 121 countries. This initially presented a hurdle in our data integration process. We needed to decide whether to limit our analysis to the 121 countries common to both data sets or find a way to incorporate the additional countries from the larger dataset. Another challenge during the ETL implementation process was regarding loading the MongoDB database and storing our transformed dataset. While thinking about the use of MySQL and MongoDB, we have decided on utilizing MongoDB as it offers more flexibility in terms of its schema-less design. Configuring the MongoDB Atlas cluster also posed additional challenges, including ensuring that the proper credentials and connection strings are successfully set up.

One of the key challenges that we face for the data analysis part is merging two data sets with a country. During the process, we noticed that one dataset contains more countries than the other, so we decided to merge the two datasets by the country names but only keep the country with the fewer countries names. The merging process also requires a clear understanding of the schema alignment to ensure the columns matched appropriately and the data retained its contextual meaning. This involved identifying the common key, country, and ensuring that it was free of duplicates, null values, or other anomalies that could compromise the merge. Choosing the appropriate join type was another critical decision, inner join was selected to focus on overlapping records, ensuring that only countries appearing in both datasets were retained.

Another hurdle that we faced during this project was coordinating our team efforts and scheduling meetings. Our project timeline coincided with the end of the semester, which meant we were all juggling multiple priorities. With final exams approaching, each team member had a full plate of deadlines and increased study loads. To complicate matters further, Thanksgiving break fell right in the middle of our project timeline. Fortunately, we managed to get together for one crucial in-person meeting just before Thanksgiving break. This meeting proved invaluable, allowing us to make significant progress on the project. We hammered out key decisions, divided tasks, and set a clear direction for our work. This boosted our productivity and team morale, highlighting the importance of such meetings. During Thanksgiving break, we faced the challenge of continuing our work while being physically apart. Knowing that we had to present our project the week we returned from break, we couldn't afford to completely pause our efforts. This meant making small edits and refinements to our work, even as we were trying to enjoy the holiday with our families. It was challenging to find the balance between project work and holiday time, but we managed to keep the momentum going through brief online check-ins and individual contributions.

Building on what we have developed in the data project 1, setting up the fundamental ETL process, the final project has given us another chance to comprehensively examine the datasets of interest with the skills and techniques we learned in the later semester. We believe that we have gained a deeper understanding of initiating the data-driven analysis process concerning the datasets that could be potentially useful for revealing certain insights on societal trends. Those skills and experiences could benefit us in the future for analyzing data under more complex scenarios.

One of the challenges we encountered during the cloud storage setup was selecting the appropriate configurations to meet the project's requirements. With several options available, such as location type, storage class, and access control, we needed to ensure that the setting was aligned with the goals of the project, given that we need secure, efficient, and accessible data storage. In the first place, the numerous choices and the technical terms of each option felt overwhelming. To address this, we conducted thorough online research, exploring Google Cloud documentation and tutorials to understand the purpose and significance of each setting option. This process not only helped us decide on a multi-region setup with a Standard storage class and uniform access control but also deepened our understanding of these technical options. Through this challenge, we developed essential research skills and knowledge about Cloud Storage, learning how to navigate technical documentation and evaluate solutions effectively. We also gained hands-on experience in cloud technologies, understanding the importance of settings like encryption and access control in ensuring data security and ease of management. These skills are valuable not only for this project but also for future work involving cloud-based tools and infrastructure. The challenge turned into an opportunity to grow our technical knowledge and confidence in problem-solving.