For our project we have chosen two data sets to analyze, Global terrorism data from the University of Maryland and historical stock performance and pricing data from S&P. The global terrorism dataset provides information on the frequency, causes, and dates of terrorist acts. By analyzing this dataset we can discover trends regarding the occurrence of terrorist attacks, the potential motivations, and the periods when these activities were more or less frequent. The stock prices and performance dataset offers a detailed view of trends in the stock market over time, including price fluctuations. Combining these two datasets presents an opportunity to explore the correlations between terrorism and international financial markets. Specifically, we will attempt to find how terrorist events across the globe impact the stock market and whether any significant stock market changes or trends relate to the motivations or frequency of terrorist attacks. One method for this is to examine different periods of heightened terrorist activity alongside stock market trends to identify any correlations.

Both of our datasets came in CSV format. They initially had a few issues that we had to clean up to use for our ETL pipelines, such as junk characters and empty rows/columns. The stock dataset was sorted by date, showing how the stock price changed from the most recent date recorded to 1970. It shows the price at open, the high, the close, and the percent change from the previous day. Our terrorist dataset, on the other hand, was more complex. It had event IDs, country IDs, attackIDs, and many other columns to help sort between the different events that occurred. It was also sorted by date, with some columns standardized, and some columns allowing for strings of variable length. It provided data from dozens of different categories, including the country the attack took place, weapons

used, gangs involved, etc. The reason we chose these specific terrorist and stock market datasets was because of the frequency and detail of the reports. For the stock dataset, we needed to pick a dataset that was representative of the entire market. The S&P 500 is widely regarded as the best way to measure market performance as a very steady and consistent indicator of how the average person's portfolio will change. Additionally, the dataset provides monthly data for several decades, which very few other datasets provide. It allowed us to get more snapshots into the way the market performance was affected by the terrorist dataset. As for the terrorist dataset, we wanted to be able to gather as much information as we possibly could about different terrorist attacks. We were not sure exactly what metrics may affect the market, and we needed to account for the possibility that metrics we thought could be predictive may have no effect at all.

One of the most challenging aspects of this project was the process of cleaning the datasets using pandas. The global terrorism dataset contained a large number of null values that needed to be handled carefully to maintain the integrity of the analysis. Removing these null values without compromising the dataset too much required a balance between preserving data accuracy and avoiding unnecessary data loss. Additionally, we encountered inconsistencies in data types within rows and had to compare variable names to their actual meanings to ensure our transformations were logical. To address this, we consulted online documentation extensively to understand how the variables were mapped from language we could understand to numeric values. This step was crucial in creating a clean and usable dataset.

Another significant hurdle was integrating Python with Google Cloud. This required setting up authentication keys, initializing the project properly, and ensuring we had the necessary Python libraries installed. Although Google provides detailed instructions, navigating these steps was complex and time-consuming, particularly for those on the team who had limited prior experience with cloud services. Once the connection was established, it enabled us to store and query our datasets efficiently, but the learning curve for this integration was steep. Lastly, writing SQL queries to connect the two datasets posed its own set of difficulties. The stock market data was organized by month, while the terrorism data contained multiple entries per month. Aligning these two datasets required us to group the terrorism data in a way that matched the monthly structure of the stock data. This grouping process was intricate, as it involved ensuring that no critical details were lost while summarizing the terrorist activities by month. Writing queries that accurately joined these datasets required a deep understanding of both the data and the SQL syntax.

One of the biggest lessons we learned during this project was the importance of clear and consistent communication, particularly around group meetings during busy periods like Thanksgiving break. Unfortunately, we did not coordinate effectively before the break, which led to scheduling issues afterward when everyone's exam schedules became a priority. This lack of communication resulted in delays and made it harder to ensure that everyone was on the same page as we moved into the final stages of the project. For future projects, we would plan to establish a clear meeting schedule ahead of time, even during holiday breaks, and check in regularly to confirm progress. This will help us avoid

unnecessary confusion and keep the workflow consistent, especially during times when individual availability might be limited.

Throughout this project, we gained valuable skills in data cleaning, ETL processes, and integrating cloud storage solutions. Using pandas to clean and transform datasets improved our technical proficiency, while setting up and authenticating Google Cloud services expanded our understanding of cloud-based workflows. Writing SQL queries to merge complex datasets enhanced our data analysis skills. However, there are still areas for further development, such as improving our time management during busy periods and deepening our knowledge of advanced SQL techniques to handle more intricate data relationships efficiently. These improvements would better prepare us for future collaborative projects.