Reflection Report – DS2002 Data Project 1

Aim:

At the outset of this project, my goal was to explore the relationship between cryptocurrency prices and the Fear & Greed Index, a sentiment-based indicator designed to capture emotional extremes in the crypto market. I was curious to see whether market emotions, specifically fear and greed, correlated with actual price movements or volatility in assets like Ethereum (ETH), Dogecoin (DOGE), and Solana (SOL).

However, as I progressed, this initial idea sparked a broader curiosity: Do external, fundamental factors also influence crypto markets in a measurable way? This led me to extend the analysis by incorporating macroeconomic indicators, specifically, Gold (GLD), Oil (USO), and the S&P 500 index (SPY), to evaluate whether cryptocurrency markets respond to broader financial trends beyond sentiment.

To evaluate these relationships, I engineered multiple correlation tests based on both price-based and sentiment-based metrics. These include:

- **Fear & Greed Index vs. Price Change**
  Explores: Does price tend to increase or decrease depending on overall market sentiment?

- **Lagged Sentiment vs. Price Change / Volatility**
  Explores: Do changes in sentiment one day influence crypto the following day?

- **Sentiment % Change vs. Price Change or Volatility** Explores: Do sudden shifts in sentiment (e.g., fear → greed) lead to higher price movement or volatility?

- **Price Change vs. Macro ETF Movement**
  Explores: Does crypto price correlate with macroeconomic forces such as gold, oil, or the broader stock market?

By structuring the analysis this way, I was able to investigate both immediate and delayed effects of sentiment, as well as test whether cryptocurrencies reflect broader financial market behaviors or operate more independently.

Challenges:

One of the primary challenges during implementation was handling and aligning datasets with mismatched time resolutions. The cryptocurrency data sourced from Kaggle was recorded at a minute or hourly level, while the Fear & Greed Index and Yahoo Finance API data were reported daily. Early attempts to merge these datasets failed or produced extremely small sample sizes (as few as 10–15 matching rows), which skewed correlation values and made the results unreliable. The solution involved aggregating crypto data to daily resolution using Pandas' groupby() and .agg() functions, and flooring timestamps with .dt.floor("d") to ensure consistency across all data sources. This allowed for successful inner joins and time-aware merges.

Another technical challenge involved deciding how to represent sentiment and price movement for correlation. While the raw sentiment score is useful, I realized that tracking percentage changes in sentiment, and comparing those to price volatility, offered a more dynamic measure of market reaction. I also experimented with lagged sentiment variables, shifting the sentiment index by one day to evaluate whether today's sentiment predicts tomorrow's price action.

Aspect That Was Easier Than Expected:

What turned out to be easier than expected was integrating APIs and handling data storage. Both the Fear & Greed Index and Yahoo Finance APIs returned well-structured JSON data, which was easy to convert into Pandas DataFrames. Saving outputs in multiple formats (CSV, JSON, SQLite) was also straightforward thanks to built-in Pandas functions like .to_csv(), .to_json(), and .to_sql().

Aspects That Were More Difficult Than Expected:

On the other hand, designing the right merge strategy for macroeconomic comparisons required more experimentation. Using a strict inner merge would have missed relevant macro data if dates didn't align perfectly. To solve this, I used pd.merge_asof() with a one-day tolerance to match each crypto datapoint with the closest earlier macro datapoint, which help improved the robustness of the correlation analysis.

How A Utility Like This Could Be Useful For Future Data Projects:

From a broader perspective, this utility has valuable applications in financial data science, trading research, and economic behavior modeling. The same ETL pipeline structure can be extended to monitor real-time data, build predictive models, or support backtesting of trading strategies based on sentiment or macro trends. It also provides a reusable foundation for ingesting, cleaning, analyzing, and storing time series data from heterogeneous sources, a common need in many real-world data science projects.