Data Science Assignment Report

Candidate: ds_Neeraj

❖ Executive Summary: Data Quality Assessment

The objective of this assignment was to analyse the relationship between trader behaviour and market sentiment, based on two provided datasets.

The first step in any data science workflow is data validation and cleaning.

Upon loading and preprocessing the data, a critical incompatibility between the two datasets was discovered. This issue prevents the core analysis from being performed.

❖ Key Findings: Data Validation Blockers

1. **Data Date Mismatch (Fatal Error)** The primary goal is to link trader actions to market sentiment on a given day. This requires the two datasets to cover the same time period.

- The **Historical Trader Data (historical_data.csv)** contains data from late 2024 to early 2025 (e.g., dates like '02-12-2024').
- The **Bitcoin Market Sentiment Dataset (fear_greed_index.csv)** contains data exclusively from 2018 (e.g., dates like '2018-02-01').

There are **zero** overlapping dates between the two files. As a result, no trades can be matched to a sentiment classification, and the merged DataFrame is empty.

2. **Missing 'leverage' Data** The assignment overview requires an analysis of trader risk, specifically mentioning 'leverage' as a key column.

- The historical_data.csv file **does not contain a 'leverage' column** or any obvious equivalent.

❖ Conclusion & Next Steps

The accompanying Python script in **Notebook_1.ipynb** is a complete and robust pipeline for the required analysis. It successfully loads, cleans, and attempts to merge the data. It includes a diagnostic check that prints the date mismatch and robustly handles the missing 'leverage' column.

As a result of the data mismatch, the aggregated DataFrames are empty. The four required visualizations (PnL, Volume, Leverage, Side) have been saved to the **outputs/** folder, but they are blank, as they are plots of empty data.

This analysis is complete to the extent possible with the provided files. The script is ready to run and will produce the required insights as soon as compatible datasets (i.e., from the same time period and including the 'leverage' column) are provided.