Domain Identification

Dataset 1: Etherium to USD Daily Price Data from 01/01/2018 - 01/02/2024

**Source**: Self-sourced from group member via TradingView

**Frequency of updates**: As per request from group members. Can be updated daily

**Metadata**: The following columns are found in the data: opening daily price, highest price of the day, lowest price of the day, closing price of the day

**Data usage terms**: Free use as the data is sourced primarily from a group member and we have obtained consent from the group member to use the data

**Rationale**: The dataset aligns with the criteria specified in the “Domain Definition” assignment; it possesses all the necessary components for our project in the finance sector. The dataset represents the daily updates in price for Ethereum. Ownership of the dataset by a group member ensures quality and accuracy as it has been directly exported from their trading software. Furthermore, the ability to update the dataset upon request and the explicit consent for use we obtained strongly support our decision to use this dataset for this project. Using the daily price data may be more feasible depending on the direction the project takes.

Dataset 2: Etherium to USD Hourly Price Data from 01/01/2018 - 01/02/2024

**Source**: Self-sourced from group member via TradingView

**Frequency of updates**: As per request from group members. Can be updated hourly

**Metadata**: The following columns are present in the data: opening price, highest price, lowest price, closing price

**Data usage terms**: Free use as the data is sourced primarily from a group member and we have obtained consent from the group member to use the data

**Rationale**: Like dataset 1 this dataset aligns with the criteria from domain definition. Like dataset 1 it shows the opening (the price at the beginning of the measurement period), the high (the highest price in the time period), the low (the lowest price of the time period), and the closing price (the closing price in the time period). We have found that these columns are sufficient to employ a project pertaining to LLMs. Furthermore, as the dataset is sourced from one of our group members we may update it as necessary and we received consent to utilize the data in any way. Using the hourly data may provide more granularity than the daily data and thus depending on the direction the project takes this may be preferred.

Dataset 3: MAESTRO

**Source**: <https://magenta.tensorflow.org/datasets/maestro>

**Frequency**: Does not seem to still be updated but there is potential for future updates as there has already been 3 versions released.

**Metadata**: Canonical\_composer:Composer of the piece. Canonical\_title:Title of the piece. Split:Suggested train/validation/test split. Year:Year of performance. Midi\_filename:MIDI filename. Audio\_filename:WAV filename. Duration:Duration in seconds.

Data usage terms: [Creative Commons Attribution Non-Commercial Share-Alike 4.0 (CC BY-NC-SA 4.0) license](https://creativecommons.org/licenses/by-nc-sa/4.0/).

Rationale: The MAESTRO dataset (MIDI and Audio Edited for Synchronous TRacks and Organization) can be used to create a model to understand and possibly even generate piano music. It comprises over 200 hours of virtuosic piano performances, aligned with MIDI data that includes key strike velocities and sustained pedal positions. The idea behind this dataset is that the model can learn from the many performances in the dataset to be able to mimic the complexity and expressiveness of human play. The only caveat is that this may be computationally expensive.

Dataset 4: ADL Piano MIDI

**Source**: <https://github.com/lucasnfe/adl-piano-midi>

**Frequency**: Does not seem to still be updated as the last change to the git repository was 4 years ago. They do state that they plan to keep updating the dataset however.

**Metadata**: Not much information found on this: Looking at the dataset, there is a zip full of midi audio files that are separated by genre. In each genre, there are subgenres before specifying specific artists. Each midi file contains its respective metadata such as name, time, and length.

**Data Usage terms**: Seems to be free to use as long as it is cited correctly.

**Rationale**: The ADL Piano MIDI dataset, in combination with the MAESTRO dataset, can be used to train a model that comprehensively understands and potentially generates music, particularly in MIDI form. By combining these datasets, the model can be trained on a diverse range of piano music, including individual pieces and extended performances, enabling it to capture various styles, techniques, and nuances present in piano music. The goal of this model would be that it can develop an understanding of musical structure, phrasing, and stylistic elements, enabling it to generate new pieces, however this may or may not be achievable with our time and resources.