Project work 1: Time Series Analysis of BTC/USD Exchange Rate

Alessia Tani, Iiro Vendelin, Sara Zambetti

Introduction

This project presents an exploratory study of the BTC/USD exchange rate. The analysis describes the series' temporal behaviour, breaks it into interpretable parts—trend, seasonality, and residual variation—and examines the autocorrelation of returns to guide model choice. It also sets up a validation scheme that preserves time order for fair evaluation. Together, these steps provide the groundwork for building and testing forecasting models in the next phases.

Data Description and Preprocessing

The dataset consists of minute-by-minute BTC/USD closing prices collected from cryptocurrency exchange feeds. In total it includes 7,278,877 observations, covering the period from January 1, 2012, 10:01 UTC to November 3, 2025, 23:57 UTC, representing almost fourteen years of continuous trading.

To prepare the series for analysis, raw UNIX timestamps were converted to UTC datetimes. The table was ordered chronologically, duplicate rows were removed, and missing entries eliminated. The spacing between observations was then examined: although the median step is 1 minute, the series is not perfectly regular (isRegular = 0), reflecting occasional gaps and uneven sampling. To place the data on a uniform grid, the series was retimed to consistent one-minute intervals, with linear interpolation applied where needed. Finally, to balance resolution with computational cost, the series was downsampled to roughly 30,000 points for plotting and aggregated to hourly frequency for the decomposition and validation stages.

Time Series Visualization and EDA

Figure 1 gives a compact view of the data. In the top panel, the full price series is shown together with a 7-day moving average. Prices clearly are not stationary: long rises alternate with drops, and swings get larger as years go by.

The middle panel shows log-returns, i.e., the change in the logarithm of price from one time to the next. This is a standard way to approximate percentage changes. In this form the series moves around zero, which is closer to stationarity. However, the size of the moves is not constant: calm periods are interrupted by bursts of big changes. This pattern is the usual "volatility clustering" seen in financial data.

The bottom panel zooms in on the most recent days, where short-term ups and downs and quick changes in volatility are easier to see. Overall, these plots suggest modeling the *returns* rather than the raw prices, while allowing for time-varying volatility and occasional large moves.

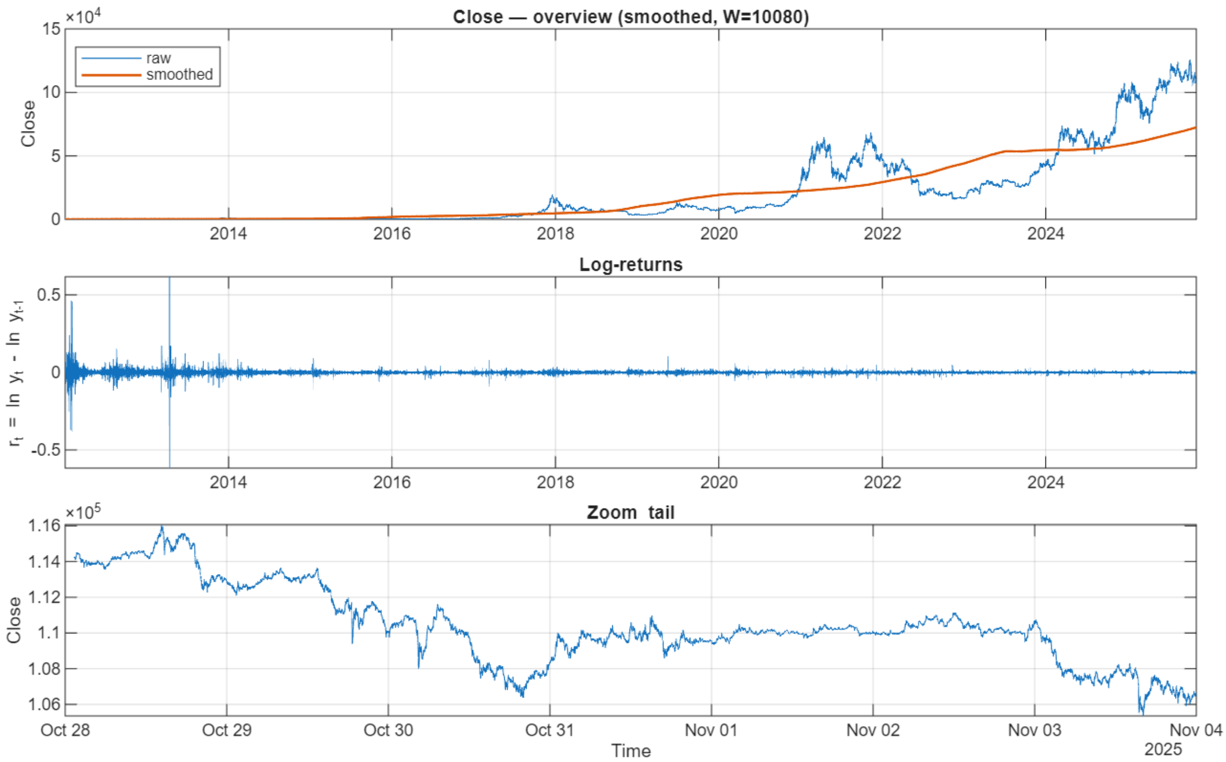

Figure 1: Exploratory visualization of BTC/USD time series.

Time Series Decomposition

Figure 2 shows how the hourly BTC/USD series over the last 12 months can be split into three parts. The data are first aggregated to hourly frequency (8,760 points) and a weekly cycle of $P = 168$ hours is used.
The observed series y_t is written as

$$y_t = T_t + S_t + R_t, \tag{1}$$

where T_t is the trend, S_t is the seasonal pattern, and R_t is the residual.
The trend traces the medium-to-long-run direction of prices. It smooths out short-term noise and makes clear when the market is broadly rising, falling, or moving sideways during the year.
The seasonal part captures a regular weekly rhythm. Its size is small compared with the price level, but the repeated ups and downs across the 168-hour cycle indicate day-of-week effects that are typical of crypto trading activity.
What remains after removing trend and seasonality is the residual. This component gathers short-lived movements and shock-driven jumps—exactly the kind of unpredictable variation that follows news, liquidations, or macro events.

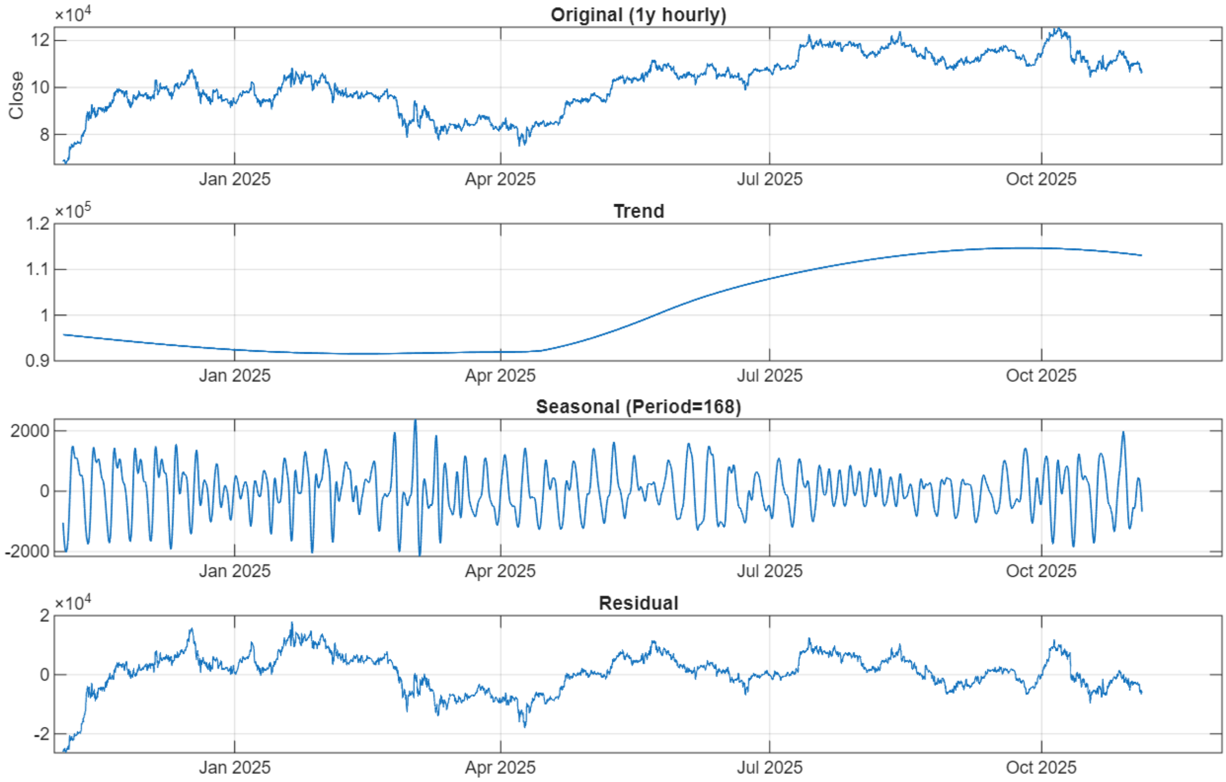

Figure 2: Hourly decomposition over the last 12 months.

Autocorrelation Analysis

The ACF in Figure 3 shows a very fast drop toward zero after the first few lags, which points to weak linear dependence in hourly returns. Almost all coefficients lie inside the usual 95% confidence bands, so the null of "no autocorrelation" cannot be rejected for most lags. In practice, this means that simple autoregressive models are unlikely to add much predictive power on the mean. There are a few isolated spikes that exceed the bands; these may reflect calendar effects (e.g., daily or weekly activity), but they do not form a clear periodic pattern. This contrasts with the decomposition on price levels, where a weekly rhythm is visible. Taken together, the evidence suggests that while prices display seasonality, return innovations are largely unpredictable from past returns alone, and models should focus more on changing volatility than on long AR dynamics.

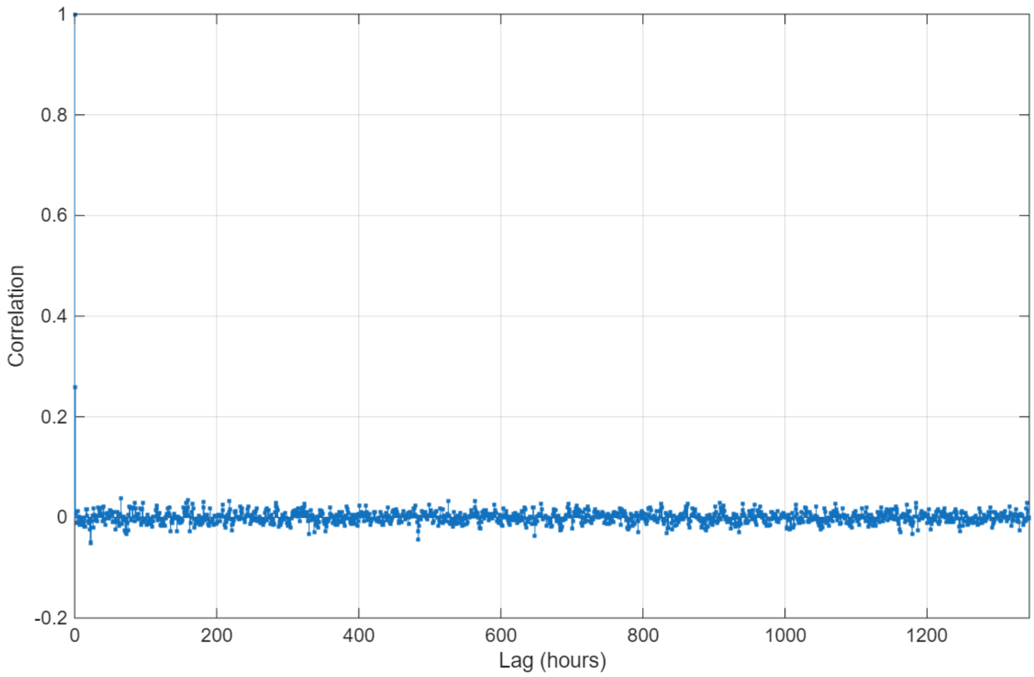

Figure 3: ACF of hourly BTC/USD log-returns up to 1,344 lags (56 days).

Data Partitioning Strategy for Model Validation

For time series, the order of observations must be preserved, so random shuffling cannot be used. To respect this temporal structure, the hourly dataset—covering the most recent 12 months with 8,760 observations—was divided using a two-stage validation strategy.

First, an 80/20 temporal split separates the data into training and testing portions. The initial 80% is used to fit and tune the models, while the remaining 20% serves as an unseen test set for final evaluation. The split point, shown by the red dashed line in Figure 4, marks the boundary between these two segments.

To improve the reliability of performance estimates, a 5-fold sliding-window cross-validation is applied within the training portion. In each iteration, the model is trained on earlier folds and validated on the following one, ensuring that validation data always occur after the training data in time. This approach provides multiple independent validation results without violating the temporal order of the series, supporting robust model selection and hyperparameter tuning.

Figure 4: Temporal split of the hourly BTC/USD data.