| FULL LEGAL NAME | **LOCATION (COUNTRY)** | **EMAIL ADDRESS** | **MARK X FOR ANY NON-CONTRIBUTING MEMBER** |
| --- | --- | --- | --- |
| Pinehas Asiedu Amankwah |  |  |  |
| Huayi TANG |  |  |  |
| Lujia Xu | Canada | Xulujia33@gmail.com |  |

|  |  |
| --- | --- |
| **Statement of integrity:** By typing the names of all group members in the text boxes below, you confirm that the assignment submitted is original work produced by the group (excluding any non-contributing members identified with an “X” above). | |
| **Team member 1** | **Pinehas Asiedu Amankwah** |
| **Team member 2** | **Huayi TANG** |
| **Team member 3** | **Lujia Xu** |

|  |
| --- |
| Use the box below to explain any attempts to reach out to a non-contributing member. Type (N/A) if all members contributed.  **Note:** You may be required to provide proof of your outreach to non-contributing members upon request. |
|  |

### Task 1 Data Quality

a. Poor quality data sample (structured)

A single price table mixes BTC-USD at ET with BTC-EUR at London time without standarizing timezone or quoto currency. It will lead to spurious returns, distorted volatilty, and invalid backtests. A normalization to UTC timestamps, and converts all quotes to a common currency is nessassary before furture actions

b. Recognization on poor data quality

In the sample above, if the timezone is not specified, I can still sanity-check alignment by taking price quotes in different quote currencies at the same recorded timestamps, coverting one into the other using the corresponding FX rate, and verifying that if that trackes the other

c. Poor quality data sample (unstructured)

Chat logs on Telegram/Discord for a new token (crypto currency). It's poor quality because the key context if missing

d. How it fails

* multiple formats and unknown time zones prevent reliable alighment with market events
* threads may collapsed, duplicated, and lack authoritative references.
* Irrelavent contents include typos, emojis make it harder to analyze

### Task 2 Yield Curve Modeling

**Model Comparision**

Fit

* Cubic Spline model is constrcted to pass exactly through all observed points, so it's perfect fit (RMSE/MAE = 0 is within expectation)
* Nelson-Siegel also has very small error. Slightly less tight in sample, but smoother and more robust than cubic spline

Interpretaion

* Nelson-Siegel's parameters are economic meaniful

1. beta0=0.024020 is the level parameter indicates a long-term rate
2. beta1=-0.009813 less than 0 is the slope parameter indicates a upward-sloping curve
3. beta2=-0.012561 less than 0 means a mild hump
4. and lambda=2.386007 is the decay scale

* Cubic spline has many cubic coefficients with no direct economic meaning. It's good for plotting but weak for factor/style interpretation

**Data Smoothing ethical issue：**  
It's really depend on the intent and transparency. Simply fitting a reproducible curve to data for summary purpose isn't usually unethical. However, it would be unethic to mislead stakeholders by hiding inconvenient data - such as kinks, periods of illiquidity, or stress, or by failing to disclose that raw data points have been dropped. Furthermore, if the explicit purpose is to misrepresent the information for dishonest reasons, it should be absolutely prohibited

### Task 3 Exploiting Correlation

### Task 4 Empirical Analysis of ETFs