Stock market prediction

1. Please see the attached dataset, data.csv, provided in csv format, which lists daily signal and price values of S&P 500.

2. The aim of this project is to test a data source (signal, second column in data.csv) which claims to be predictive of future returns of the SP500 index (spy_close_price, third column in data.csv). We use SPY (SPDR S&P 500 ETF) as a proxy for the SP500 index.

3. The signal and spy_close_price are both received at the same time at the end of the day on the date listed in column 1. We do not know how the signal is generated or have a prior conviction about the forecast horizon over which the signal is supposed to be effective, nor its stationarity.

4. The first step in this endeavor is data cleaning. Assume all values in data.csv are potentially suspect, and please identify any errors in the data, flag them with a note, and suggest a corrected value or if advisable, you may choose to ignore them for purposes of your analysis. Please explain what types of analysis you did to identify the errors, and provide any assumptions/intuition/formulas/scripts you may have used to help you find them.

5. Given the cleaned/censored version of the data you created in (4), please perform an analysis of the predictive power of signal with respect to spy_close_price. This analysis could take various forms ranging from qualitative, to linear regression to recurrent neural networks, and everything in between. Feel free to use whatever technique(s) you feel are most appropriate, taking into consideration: a) your general familiarity, b) their potential for success on this task, c) the time involved. We believe a critical factor distinguishing great researchers, besides their mastery of a field, is their ability to prioritize what to work on. Use this as an opportunity to demonstrate that ability. Similarly, just as in research, negative results are nothing to be ashamed of (and often quite informative), so please share all your ideas/attempts, even if they proved less than successful (of course, a guess as to why they didn't work or how to improve them would be great as well).

Please document the experiment(s) you performed (including relevant code, package references, etc) and summarize your conclusions about the viability and shortcomings of this signal as a predictor of spy_close_price, including any materials you feel are appropriate to support your conclusions (eg, graphs, tables, etc). Use jupyter notebook. If there were other experiments you didn't have time to perform, or future avenues of work you might like to pursue, please discuss those as well (we may work on these ideas together as a follow-up).