WORLDQUANT UNIVERSITY

MASTERS OF SCIENCE IN FINANCIAL ENGINEERING

MACHINE LEARNING IN FINANCE (C18-S3)

ALEXANDER BOTICA

ALEXANDER VICTOR OKHUESE

CAIO ARANHA VINCHI

DAVID KWASI NYONYO MENSAH-GBEKOR

DAVID WONDER DOE-DEKPEY

GROUP WORK PROJECT – FIRST SUBMISSION

DATA PREPARATION OVER A SET OF RAW TICK DATA INTO BARS AND STATIONARITY ANALYSIS

GROUP 2-A

2019

ABSTRACT

As data preparation is a fundamental part in implementing financial data structures, this document makes of use of 20-days raw tick data from the S&P500 E-Mini features to generate time, tick, volume, and dollar bars with the intent to verify its stability, correlation, and normality via common econometrics.

Keywords: Machine Learning; Finance; Tick Data; Bars Stability; Serial Correlation; Jarque-Bera Normality Test.

Methodology

Following the instructions explicitly provided in this course notes, this project makes use of lecturer Jacques Francois Joubert' GitHub repository code, properly cloned via the documentation available at https://github.com/Jackal08/financial-data-structures.

The required tasks are comprised of the following:

1. Create tick, volume, and dollar bars. (Bar must have open, high, low, and close values.) Students can do this from first principles or clone the following repo for an implementation.

2. Count the number of bars produced by tick, volume, and dollar bars on a weekly basis. Plot the time series of the bar count. What bar type produces the most stable weekly count? And why?

3. Compute the serial correlation of each bar type and report back on which method has the lowest serial correlation.

4. Apply the Jarque-Bera normality test on returns from the three bar types. What method achieves the lowest test statistic?

The data chosen for this project, also following the guidelines available in the repository, refers to the E-mini S&P500 futures from 1 September of 2013 to 20 September 2013 sourced from the 288mb csv file from Tick Data LLC available at: https://s3-us-west-2.amazonaws.com/tick-data-s3/downloads/ES_Sample.zip; from which internal dependencies generate the required tick, volume, and dollar bars. (More details about the specifics of each dependency for different operating system can be found in the referred repository).

The code is available in the folder of this project labeled as ML_GP_Sub_1.ipynb, accompanying html file for ease of use.

Solution Design

According to Folger (2018), data-based chart ranges allow us to view the price change of assets from various data ranges rather than time slots. Bar charts of volume, dollars, and ticks are examples of charts based on data. These charts represents a bar at the end of a specified data range, regardless of how much time has passed:

- Tick: indicate a specified number of transactions.
- Volume: indicate when a number of stocks or contracts have been traded.
- Dollars: record the value of the asset at the time a given accumulated transaction value is reached.

Tick, Volume, and Dollar Bars

The below with five entries from the generated dollar bars are used as an illustration of the transformation generated from the code through the raw data format regarding indicators captured; those include usual information such as time interval, opening, high, low, and closing prices, volume per transaction and volume-weighted average price over the trading horizon.

date	open	high	low	close	volume	vwap
9/1/2013	$1,640.25	$1,642.00	$1,639.00	$1,641.25	21722	1640.667
9/1/2013	$1,641.25	$1,643.50	$1,639.75	$1,640.75	21706	1641.83
9/2/2013	$1,640.75	$1,644.50	$1,640.50	$1,644.50	21652	1643.15
9/2/2013	$1,644.50	$1,646.00	$1,642.75	$1,645.00	21709	1644.557
9/2/2013	$1,645.00	$1,647.25	$1,644.25	$1,645.50	21646	1646.012

Table 1 – Head example for the data generated dollar bars.

Stable Weekly Count

Via inference reasoning, analyzing each bar count plotted against one another, it's not possible to plausibly make strong assumptions regarding its stability, has all of them have more or less the same count.

date	tick	volume	dollar
9/1/2013	2	2	2
9/8/2013	388	351	353
9/15/2013	325	305	312
9/22/2013	376	343	357

Table 2 – Tick, volume, and dollar bars weekly count

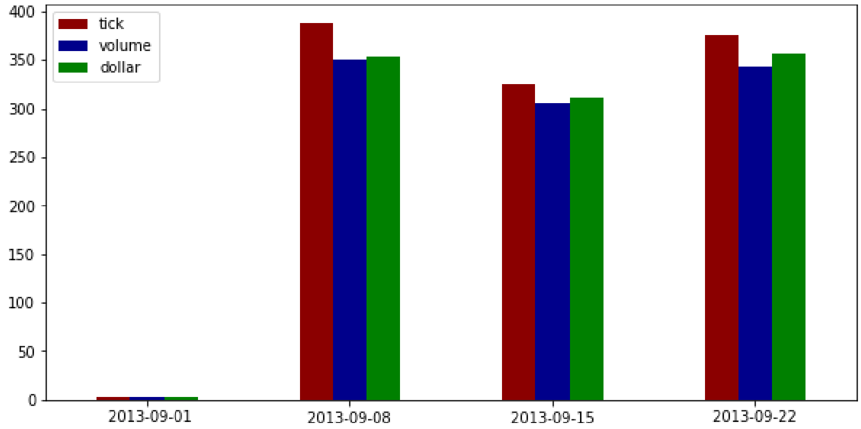

Figure 1 – Number of bars over time

After introducing coherent metrics such as the standard deviation at the 95% level, discounting the low volume of observations, it's still not possible to make strong assumptions as their divergence are empirically insignificant.

tick	47.29%
volume	47.77%
dollar	48.04%

One last attempt is made by scaling all the bar counts and plotting it into a time-series graph without success; thus confirming that given the size of the data (20 days) is not a reasonable sample size from which stability inferences can be made either via obvious inference analysis or using standard statistics.

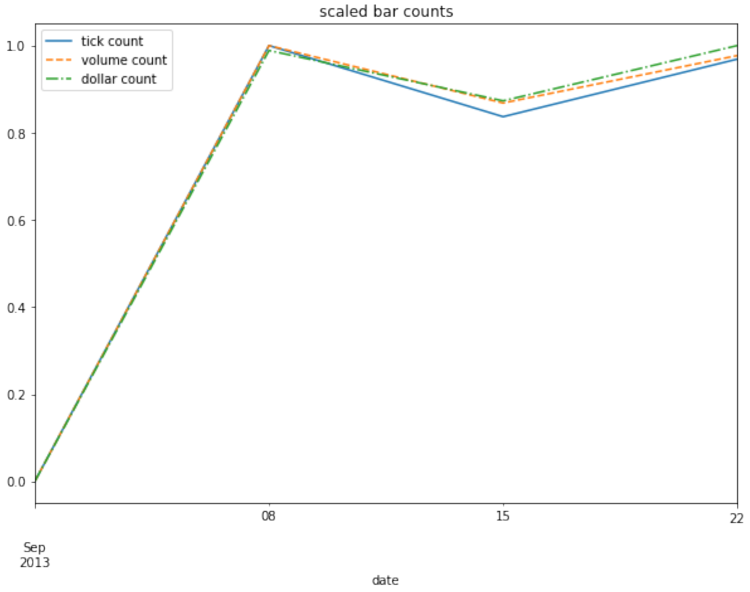

Figure 2 – Time-series scaled bar count for Tick, Volume, and Dollar Bars

Serial Correlation

For a weakly stationary process, the theoretical value of a simple autocorrelation for a particular number of lag(s) is the same across the whole series both backward and forwards.

In this case, we implement a short analysis consisting of 10 lags and plot the results as a way to analyze whether this time-series may present significant oscillating behaviour, persistence, a possible order of differencing and/or moving averages.

The computed the serial correlation for the bars are:

	sample size	statistic
tick	1090	-0.0028
volume	1000	0.0308
dollar	1023	0.0332

Table 3 – Autocorrelation for tick, volume, and dollar bars.

As expected, the autocorrelation plots from the calculated bars returns yield a similar pattern throughout all the bars, which given the sample size and inherent composition demonstrates insignificant both positive and negative shock oscillatory behavior. Without further capturing higher moments, no linear dependency between the explanatory variables is therefore detected, nor is any lag sufficiently big to justify autoregressive and/or moving average components.

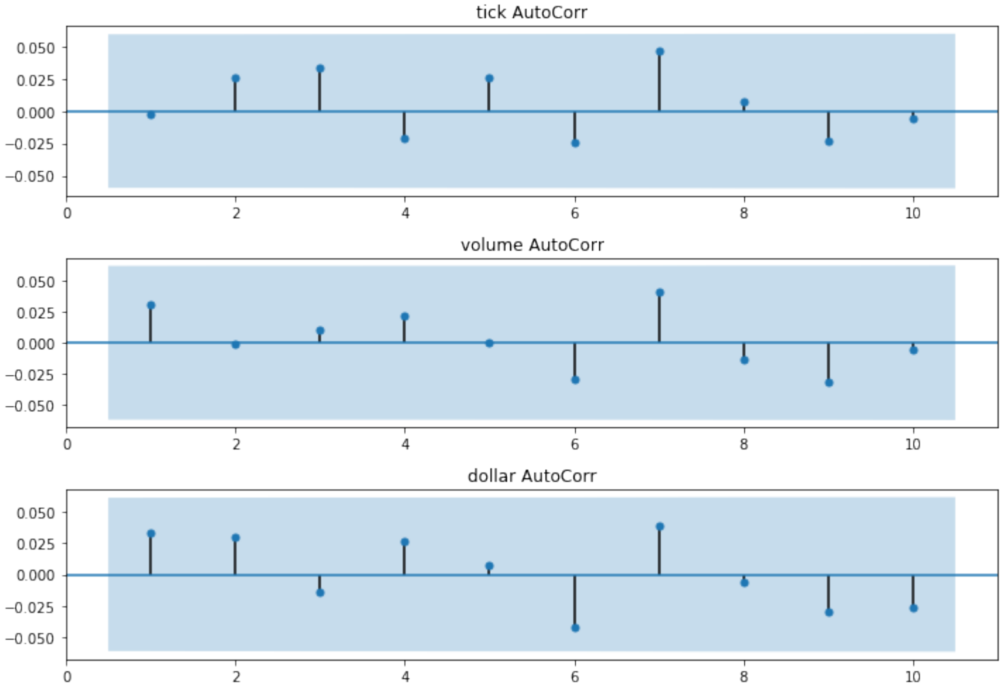

Figure 3 – Autocorrelation Function Plots for Tick, Volume, and Dollar Bars
Note: the zeroth lag has been removed from the plot.

The auxiliary histogram of the autocorrelation corroborates the assumption above, although demonstrating an apparent right behavior.

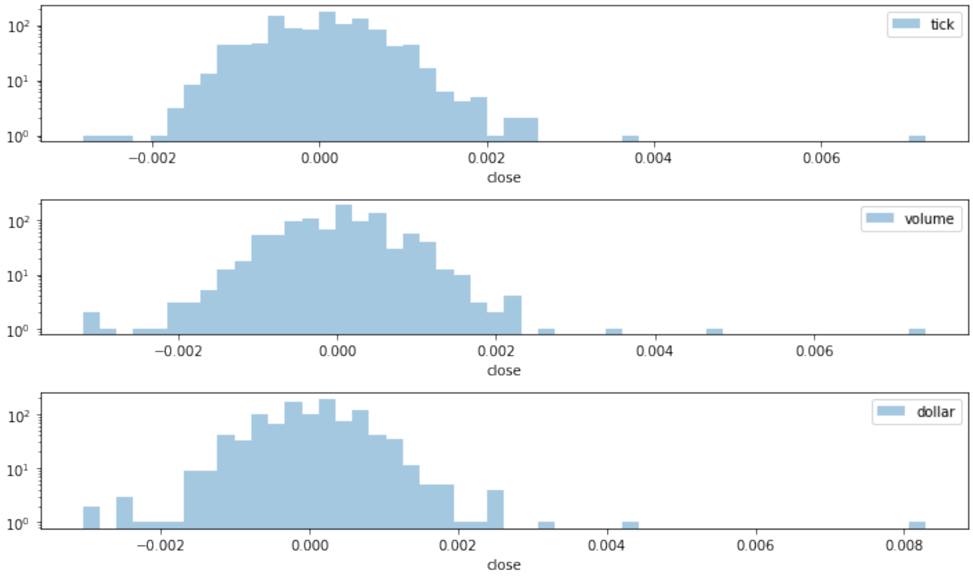

Figure 3 – Autocorrelation histogram for Tick, Volume, and Dollar Bars

Normality Tests

To test whether residuals derived from the calculated bars returns are normally distributed, this project makes use of the Jarque-Bera and Shapiro-Wil tests.

For the first, without a complete analysis of the relative components and additional criterions, nor a desirable sample size given its asymptotic nature over the Chi-Square distribution, or a relative distribution of its p-value, but just by observing its high number, it's possible to infer that errors might not be normally distributed, this is heavily accentuated in the dollar bars due to its magnitude.

For the second, although its numerical model is in favor of higher sample sizes, the yield low statistic also falls into a possible rejection of normality at the 5% confidence level.

	sample size	**statistic**
tick	1090	4522
volume	1000	4911
dollar	1023	11917

Table 4 – Jarque-Bera normality test for tick, volume, and dollar bars.

	sample size	**statistic**
tick	1090	0.9463
volume	1000	0.9314
dollar	1023	0.9120

Table 5 – Shapiro-Wilk normality test for tick, volume, and dollar bars.

To conclude, even with the small size, and without taking into account time bars, after plotting a histogram of the bars, the normality conclusion is to reject it in favor of fat tails.

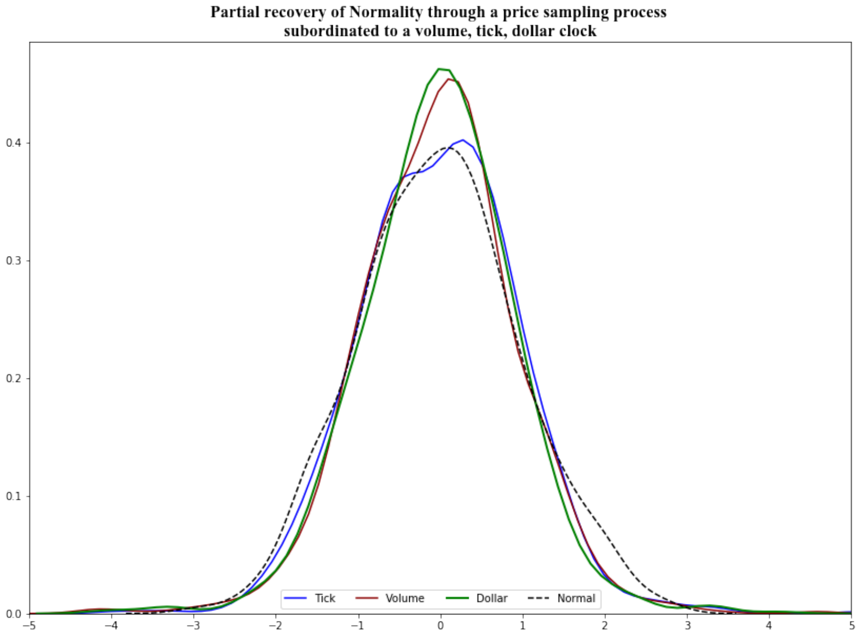

Figure 3 – Partial Recovery of Normality via Price Sampling Process

This should be enough to conclude that a much higher number of observations is desirable as to avoid missing valuable and intrinsically desirable features over risk management, including more a robust metrics approach, both in a macro perspective regarding a possible change in regime, as with smaller and smaller time intervals prone to the increase in variance (according to the Information Theory) and effects such as slippage, overlapping time zones ordering, and heavily management incision.

References

De Prado, M.L., 2018. Advances in financial machine learning. John Wiley & Sons.

Francois Joubert, J. (2018). *Cookiecutter Data Science*. [online] Drivendata.github.io. Available at: https://drivendata.github.io/cookiecutter-data-science/ (Accessed 5 Sep. 2019).

Jarque, C. M. and Bera, A. K. (1980) *Efficient Tests for Normality, Homoscedasticity and Serial Independence of Regression Residuals, Economics Letters*. Available at: https://pdfs.semanticscholar.org/8314/ae1ebfa961eb6e7ff3574c96295bdedacc7a.pdf (Accessed: 19 April 2019).

Jean Folger (2018) *Advantages of Data-Based Intraday Charts, Investopedia*. Available at: https: //www.investopedia.com/articles/trading/10/data-based-intraday-chart-intervals.asp (Accessed: 19 April 2019).