Yucehan Kucukmotor

CIS3920, Data Mining

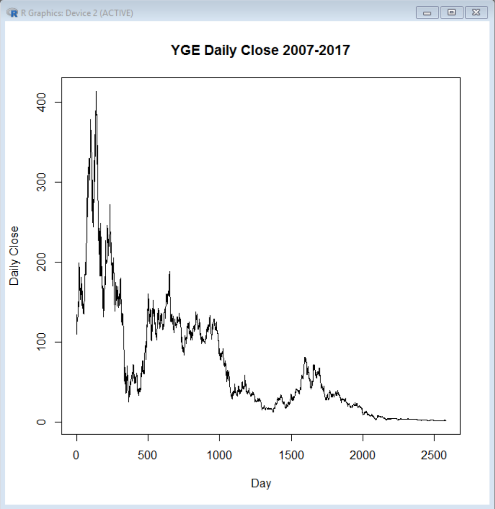
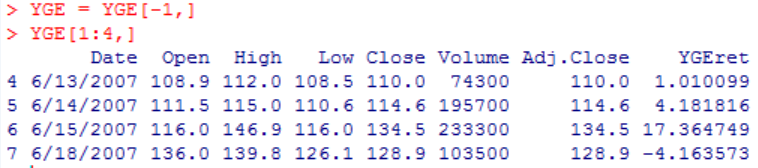
Prof. Lawrence Tatum

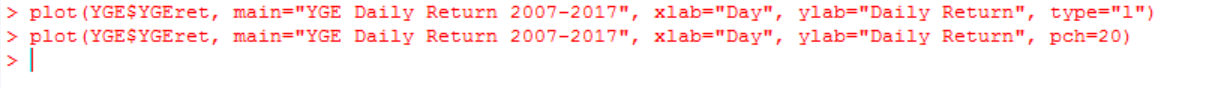
Due 09/09

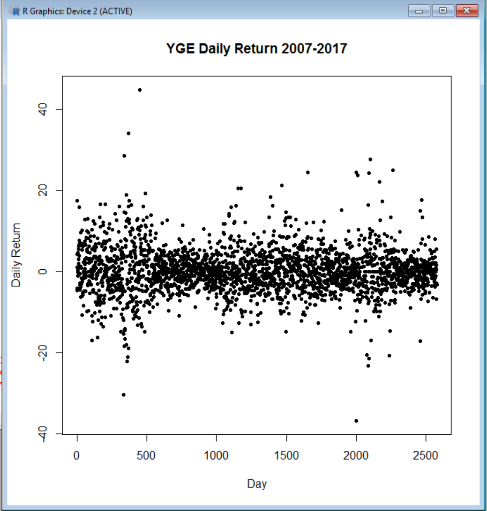
1. Company name: Yingli Green Energy Holding Company Limited.

NYSE: YGE

Interesting: Yingli Green Energy was one of the main sponsors in FIFA 2014 World Cup which was held in Brazil. Moreover, Yingli Green Energy has supplied models for stadiums such as Maracana Stadium in Rio de Janeiro. Yingli Green Energy is also a proud sponsor of FC Bayern Munchen and the U.S. Soccer Women’s and Men’s National Teams, which is very great on my end since I’ve been a huge soccer fan ever since I was a kid. Knowing the company I am interested in is also interested in sponsoring what I am interested in: soccer.

1. C:\Users\Yucehan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\2.2.pngA) Removed column that resulted in NA, and then displayed the first four rows and all the columns again to see if there was any other N/A, or any other problem for that matter, was present. Then used plot function to graph time series of daily closing prices. Graph to this is to the right. “YGE Daily Close 2007-2017”

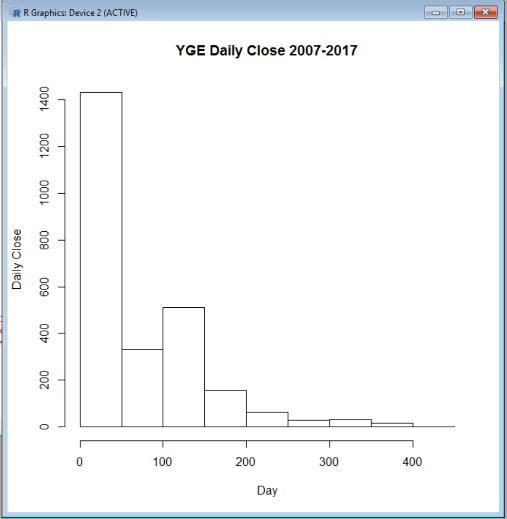
 B)

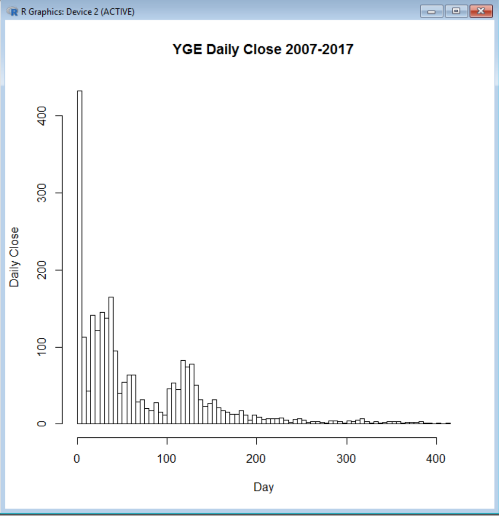
Above lines are to show how I used plot function again, but this time to show daily returns but NOT the daily closing prices. This is found using the formula: 100\*(G3-G2)/G2 where G3 corresponds to the closing price of the day and G2 corresponds to the closing price of the previous day. First line, however, was to show the mistake I made. When I used type=”l” (lowercase L), graph did not use horizontal clustered points, but lines that were not very possible to read.

Graph to the right is to show my time series plot for daily returns from 2007 to 2017.

C) Second one is more stable since there are random walks in the first graph. “There is no stable underlying population average to estimate, nor an underlying population standard deviation.” (Tatum, LN2, 8-9). To make a graph that makes sense free of random walks, we need to use percentage changes, which is the second graph I provided. It is not possible to read the future data using the first graph since it is useless. Conclusion is that second graph is more stable.

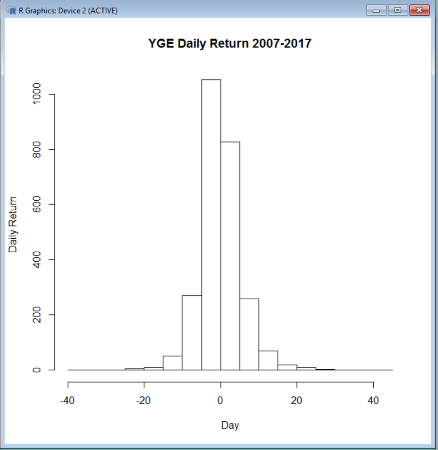
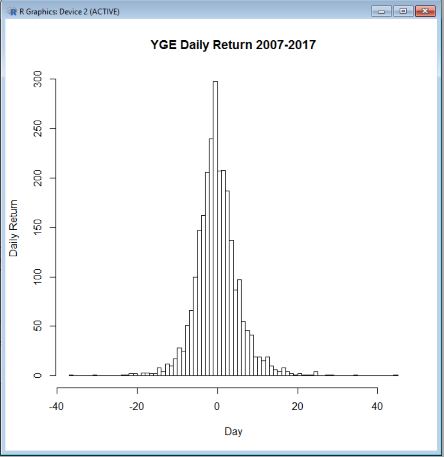
D) Yes, it is coming from a more stable source.

1. C:\Users\Yucehan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.1.pngA) Used hist function to create an R histogram of the closing prices.

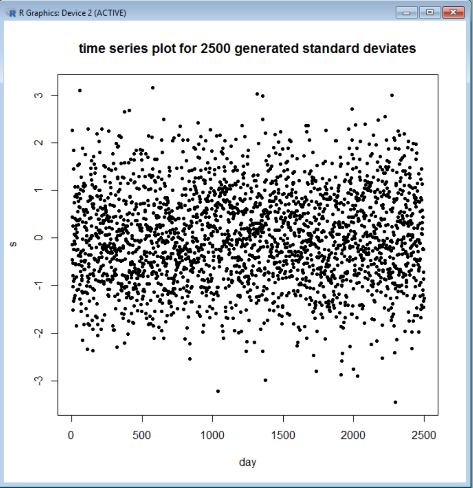
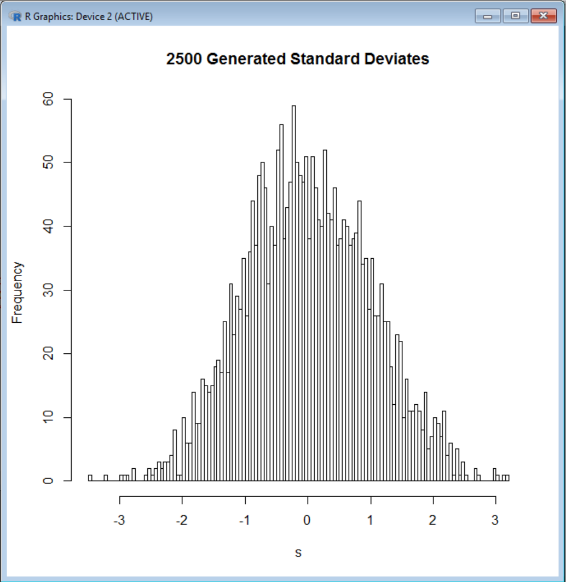


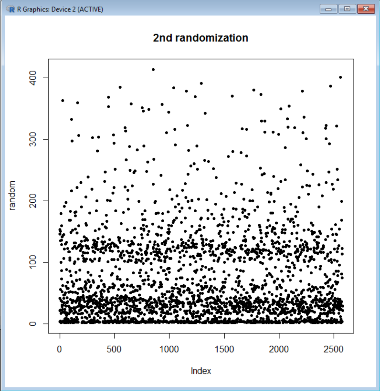
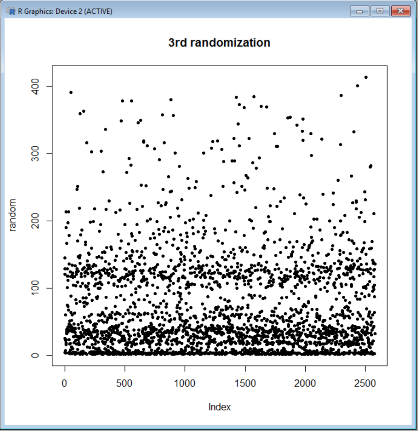
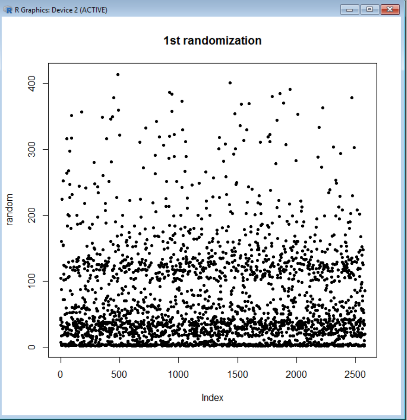
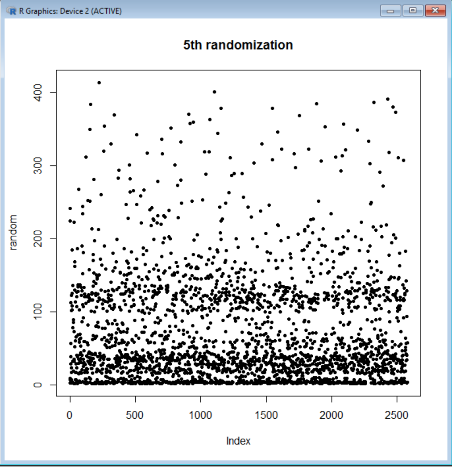
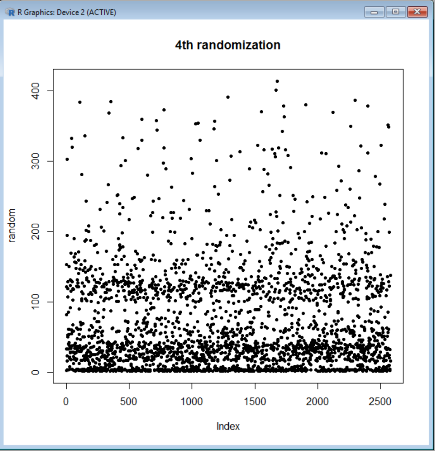
B) Used hist histogram again, this time, I have not used the default number of bins, but 100 of them instead. Lines are beneath the graphs. First line was to see how it would look when I used default instead of 100 bins.

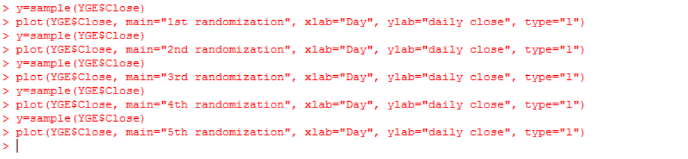
C) When we use 100 bins instead of 10, it is easier to catch a trend among the closing prices. However, it is not a good idea to use daily closing prices to build a forecast for the next closing prices.

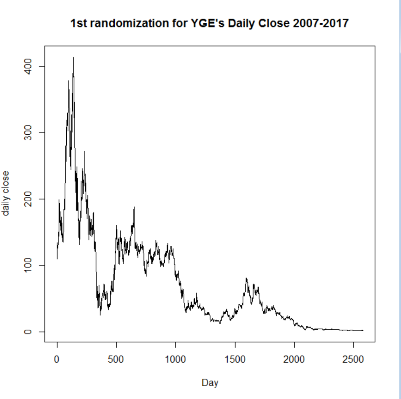
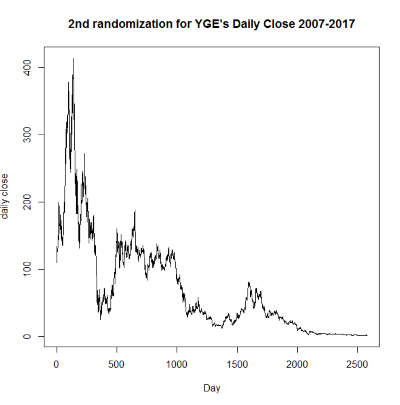
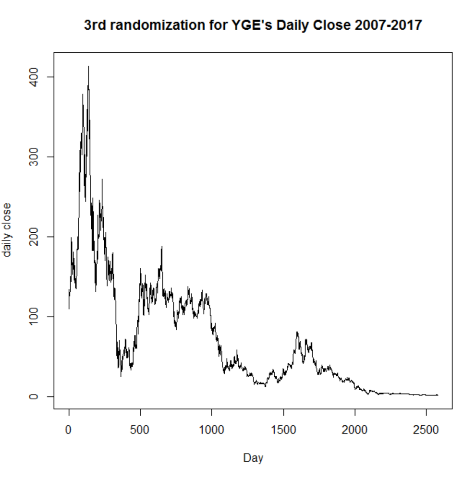
1. C:\Users\Yucehan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\4.1.1.png
2. and B) are on the right.
3. Bell curve is solid. Solid bell-curved look when we use 100 bins is the evidence of an underlying probability distribution.

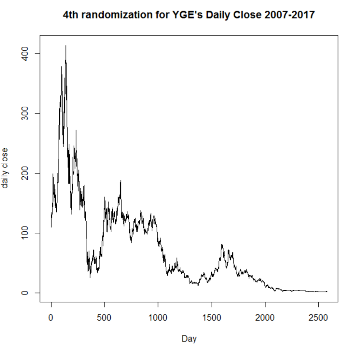
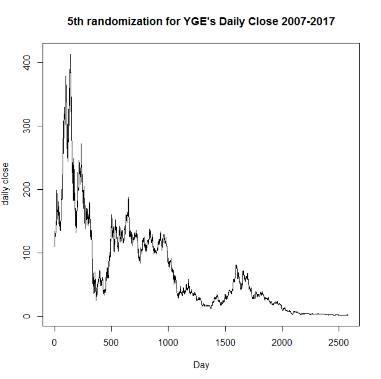
C:\Users\Yucehan\AppData\Local\Microsoft\Windows\INetCache\Content.Word\5.1.png5.

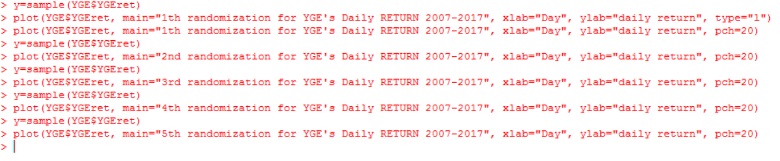
1. Graph to the right is for generated standard deviates. Points all scattered all around as numbers are generated randomly.
2. Comparing the two graphs, I see there’s a huge difference between the two time-series plots. Graph to the right (2500 generated standard deviates) is clustered around 0 line, yet it is almost impossible to read data from this. It may be a good idea for certain cases to use generated standard deviates for modeling purposes for stock returns, but I think, it is not the case here. I am not able to read the extreme values and numbers are not clustered very neatly, making the graph hard to read for logical deductions.
3. A histogram of 2500 generated standard deviates with 100 bins is to the right.
4. Yes, it seems like a good model when we compare the two histograms (hist. of returns and hist. of 2500 generated standard deviates). It is good for modeling purposes in stock returns since it makes the market easier to predict. When we have a normal distribution, it is likely that we won’t get extreme spikes in our graph. Thus, using a model by generation standard deviates help us get rid of the random spikes which may lead to a better prediction and forecast.

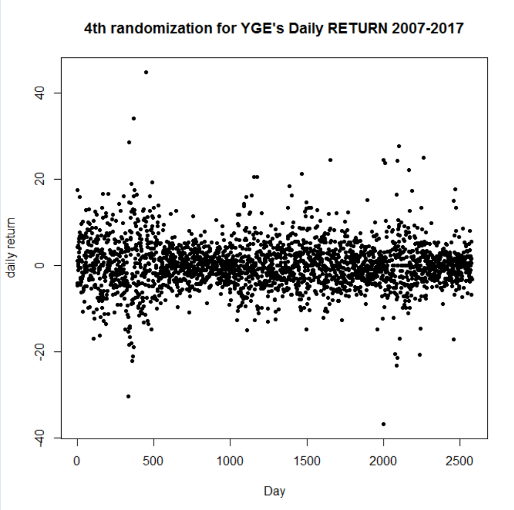
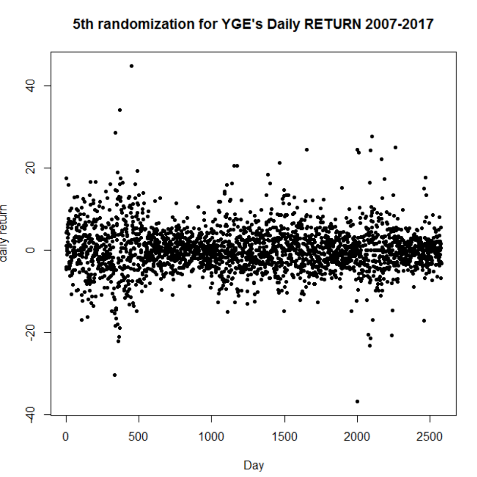
6. A) to the right is 5 graphs. Created by using the plot function and pch=20 format. All 5 randomizations look alike, although they are not the same. However, I am pretty sure I made a mistake along the way, and this is not what I should be posting here. It was useful though, since I could play around with different methods and functions and learn.

Graph to the right shows my work and how I graphed randomization using a different method (type=”l”).

 Using the graphs on the right and on the bottom, we can conclude that time-series was stable. They all look the same with the original one (one page 1).





1. Used sample function to generate randomization and used plot function to graph the daily return values. All the graphs look the same, including the original. Therefore, we can conclude that time series was stable. Below are the graphs I created. (4 and 5, they were all the same)

7) No, it would not. Because occurrence remains the same even when we randomize the numbers. Data is the same data, therefore shape of the graph must not change. Same for sample average and standard deviation. They remain the same even after randomization.