1. **To handle the data:**
   a) Converted ‘GBp’ to ‘GBP’ in the London data
   b) Removed the dates that were not there in either one of the datasets
   c) Converted both the datasets to prices in USD for that particular day for better development of the model.

   Model:

   a) Split the datasets into 60-40 train-test ratio in order to make sure the model works for fresh data
   b) The rationale behind the model is to use **Pair Trading** which is a Market Neutral Statistical Arbritrage Stategy. We see that the two stocks are of the same company, so pair trading would be useful.
   c) In order to get this working, we first check the correlation of the prices of the datasets (> 0.9).
   d) Now, the important part is that time series data of stocks is generally because of a trend, so we need to check if there is **co-integration between the stock prices**. That means, even though the stock prices are non-stationary, a linear combination of it is stationary.
   e) Given that the linear combination of the stock prices is stationary, we run a regression to find the predicted London Stock Price from the Australian one.
   f) Then we trade on the residual of the OLS regression, because we know that the **residual will oscillate around the mean**.
   g) Each time we trade, **we create a perfectly hedged trade**, with 100 London Stocks and taking the other position in the equivalent amount of Australian Stock for the day.
   h) The model assumes **no transaction cost** for trades and being able to **buy fraction number of stocks** to perfectly hedge the position

2. To validate the model, we split the data into training and test sets.
   For each set, first we check:
   a) If the **correlation** between the London and Australian stocks is high enough (> 0.9)
   b) We check with a significance level of 10%. Null hypothesis is that there is no co-integration. **We use the Engle and Granger approach built in in python to check for co-integration**. We only move ahead by rejecting the null if the p value < 0.1, since co-integration is an important factor.
   c) We check if the **OLS regression** run on the stock prices is statistically significant
      |t| > 1.65, at the 10% significance level.
      Y -> London Stock

X -> Australian Stock
We use **robust** Standard Error to account for heteroscedasticity

d) We check if the **residuals are normally distributed**. Null Hypothesis being that the residuals are normal.
We only move ahead by failing to reject the null if the p value > 0.1, since only then can we use the Z-score

```
Training set
Co integration check: True
P value: 0.000108627093269

Correlation Coeff: 0.994027035021
OLS t-statistic: 864.586593585

P-value for normality check: 0.468346327566
Residuals are normally distributed: True
```

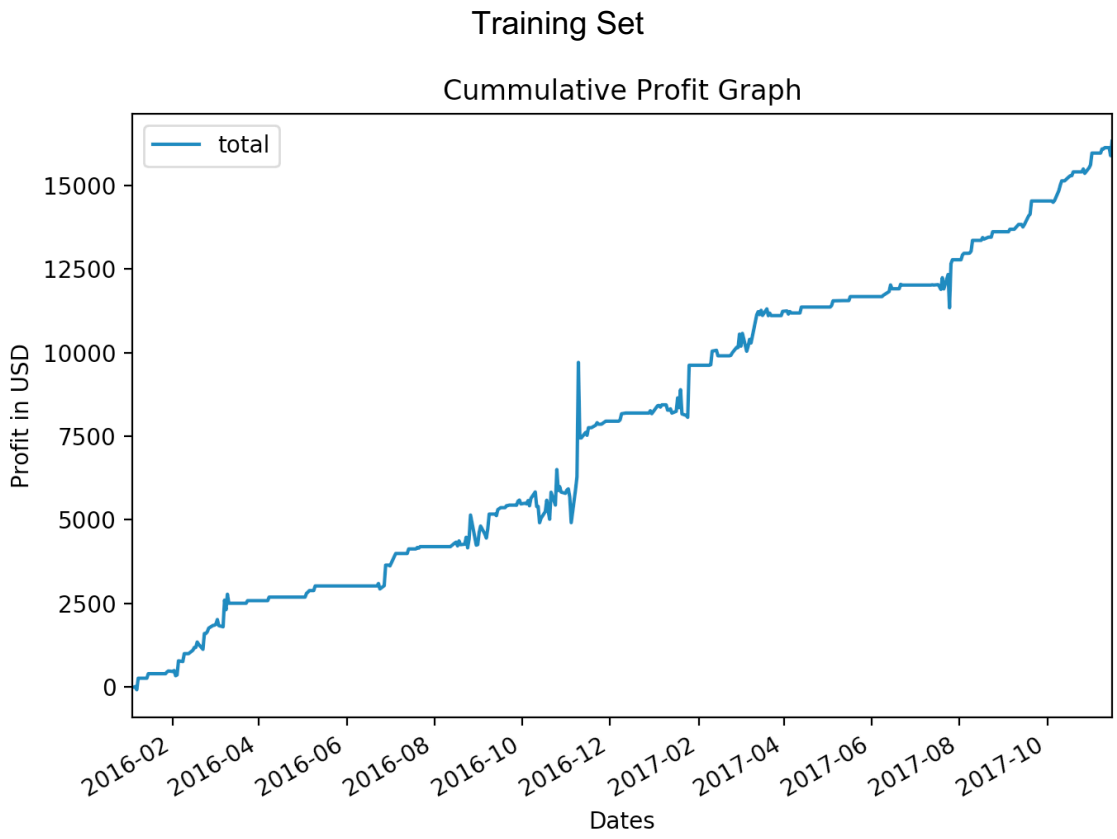
```
Test set
Co integration check: True
P value: 0.0988375144002

Correlation Coeff: 0.944560345704
OLS t-statistic: 618.300602053

P-value for normality check: 0.530123526821
Residuals are normally distributed: True
```

3. The parameters in my strategy are:
   a) **To Long London stock and Short Australian:**
   If the **residual is less than 1 standard deviation below the mean**, that means that London stock is undervalued or Australian is overvalued. Importantly, we create a perfectly hedged trade, where we long 100 London Stocks, and short equivalent ratio of Australian stocks
   b) **To Short London stock and Long Australian:**
   If the **residual is more than 1 standard deviation above the mean**, that means that London stock is overvalued or Australian is undervalued. Importantly, we create a perfectly hedged trade, where we short 100 London Stocks, and long equivalent ratio of Australian stocks
   c) **Exit the position:**
   When the residual **is between 0.25 SD of the mean**.

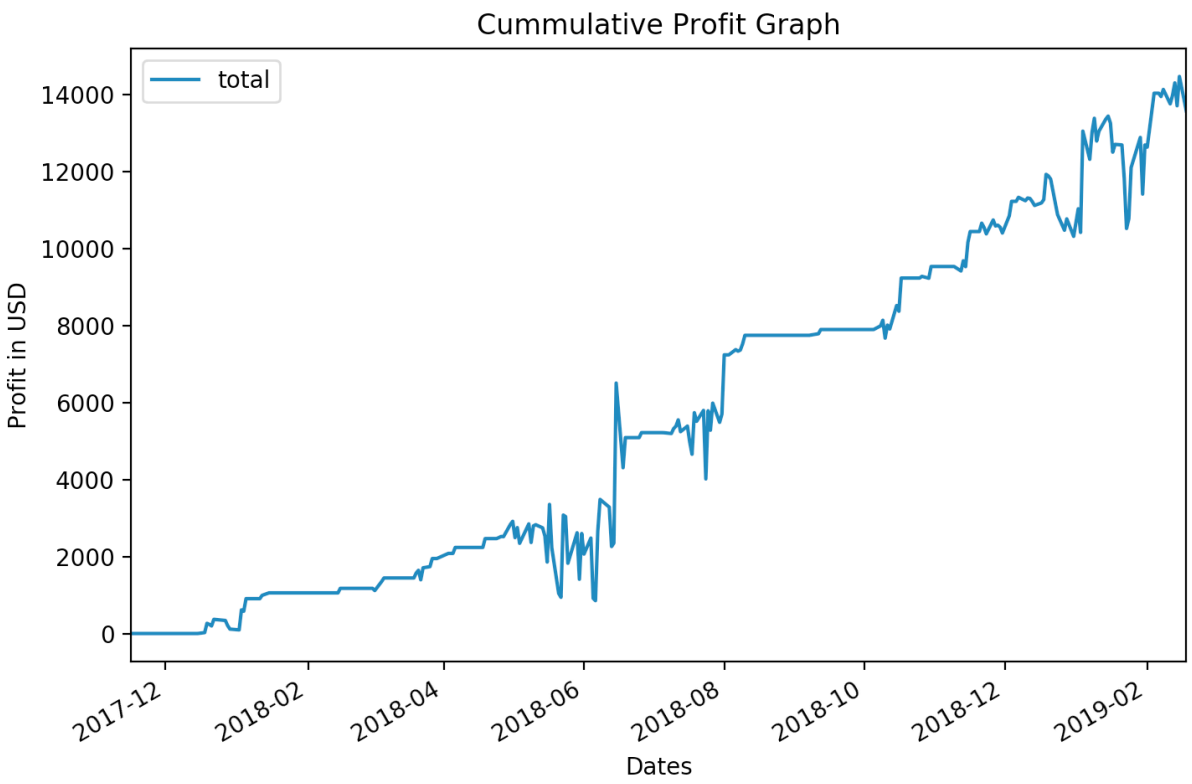
d) **Stop Loss:**
To prevent humungous losses, we exit our position if it is **more than 3 SD on either side of the mean.**

To come up with these numbers, I mainly used a graph to see how much the residual is oscillating around the mean, along with some trial and error on the training set

4. **Important Note:**
Since in the pair trading strategy I am using**, I start off with 0 capital** because for every long and short, the amount is perfectly hedged, I am not being able to gauge the return on the investment since the starting capital is 0, and thus cannot calculate the other parameters like the Sharpe Ratio etc.
I have attached the profit graph that is being generated for the strategy where according to the signals mentioned in point 3 above.



Test Set



5. **Limitations**:
   a) The model does not account for transaction cost for trades
   b) Assumption of being able to buy fraction number of stocks to perfectly hedge the position
   c) Because of small dataset, only using 10% significance level instead of 5%.

   Risks:
   a) Forex risk because the entire model is based on USD
   b) Execution Risk because we need to long and short simultaneously
   c) Australian Market opens before the London Market

6. In order to improve my model further, I would try to get more data, in order to use a better significance level to be more confident about my results.
   I would also try to get the dividend information to calculate the adjusted close price instead of the normal close price.