ZERODHA

Trading Systems

ZERODHA.COM/VARSITY

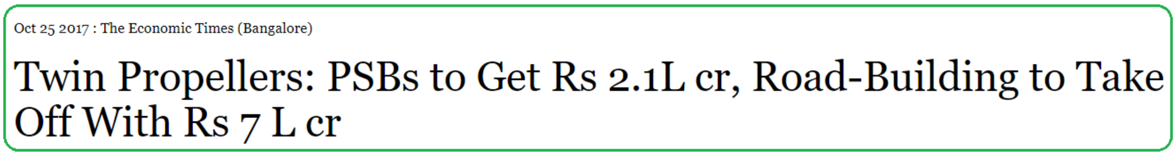
Table of Contents

CHAPTER 1

What to expect?

1.1 – What is a trading system?

Such a glorious day to start this module! Here is the headline that rocked the stock markets today –

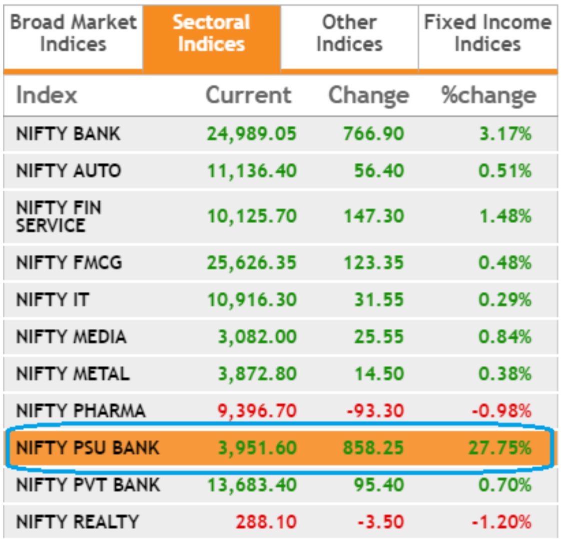


Oct 25 2017 : The Economic Times (Bangalore)

Twin Propellers: PSBs to Get Rs 2.1L cr, Road-Building to Take Off With Rs 7 L cr

Yesterday i.e. 24th Oct 2017, the Finance Minister announced that the Government would infuse Rs. 210,000 Crore into the Public Sector banking system, which is basically an effort to save the PSU banks from the deteriorating NPAs (Non-performing assets).

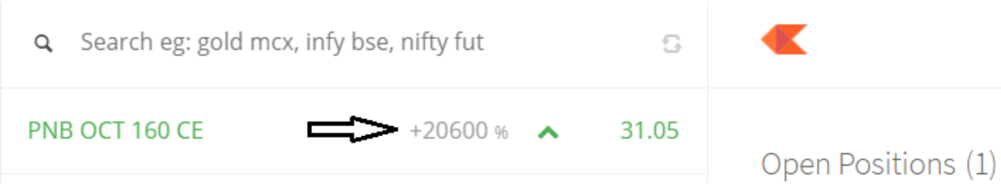
How did PSU Banks react to this announcement? After all, this is a lease of life to the PSUs. Well, they were jubilant, as expected –



Index	Current	Change	%change
NIFTY BANK	24,989.05	766.90	3.17%
NIFTY AUTO	11,136.40	56.40	0.51%
NIFTY FIN SERVICE	10,125.70	147.30	1.48%
NIFTY FMCG	25,626.35	123.35	0.48%
NIFTY IT	10,916.30	31.55	0.29%
NIFTY MEDIA	3,082.00	25.55	0.84%
NIFTY METAL	3,872.80	14.50	0.38%
NIFTY PHARMA	9,396.70	-93.30	-0.98%
NIFTY PSU BANK	3,951.60	858.25	27.75%
NIFTY PVT BANK	13,683.40	95.40	0.70%
NIFTY REALTY	288.10	-3.50	-1.20%

As you can see, the PSU Bank index shot up 27.75% at opening.

Some of the PUS stock options were on steroids, here is the hero of the day –



Search eg: gold mcx, infy bse, nifty fut

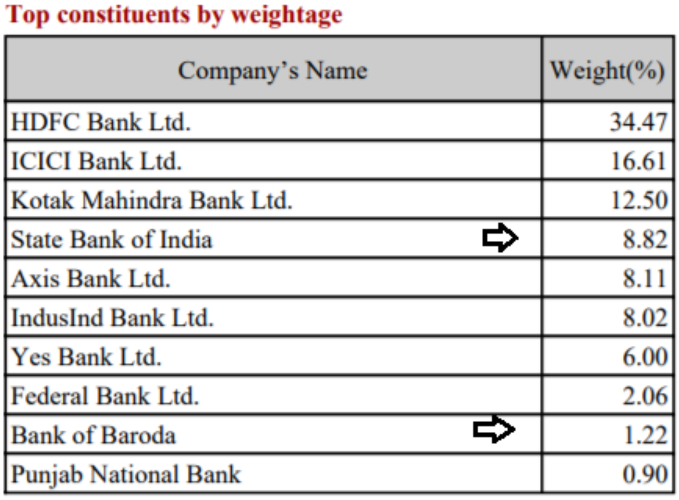
PNB OCT 160 CE +20600 % 31.05

Open Positions (1)

Punjab National Bank's 160 Call option expiring on 26th Oct 2017, shot up 20,600% overnight! If you had bought 1Lac worth of option on 24th Oct, it would have translated to 2.02 Cr on 25th Oct morning. So clearly, there is a lot of action in the market today.

Earlier in the day, my colleague and I were looking at the way markets were behaving and trying spot an opportunity, and here is something that looked interesting –

Bank Nifty Index too joined the party, with the index going up nearly 3% (look at the image of the sectoral indices above). However, a 3% move on Bank Nifty was quite questionable considering the fact that PSU banks contribute just around 10% to the Bank Nifty index, look at the index constituents and its weights below –



Top constituents by weightage

Company's Name	Weight(%)
HDFC Bank Ltd.	34.47
ICICI Bank Ltd.	16.61
Kotak Mahindra Bank Ltd.	12.50
State Bank of India ⇨	8.82
Axis Bank Ltd.	8.11
IndusInd Bank Ltd.	8.02
Yes Bank Ltd.	6.00
Federal Bank Ltd.	2.06
Bank of Baroda ⇨	1.22
Punjab National Bank	0.90

Considering this, my colleague and I decided to write a short strangle on Bank Nifty and collect a premium of close 253 points per lot, obviously hoping that the volatility would die and premiums would reduce.

I don't want to debate about the reasoning of this trade – whether it's going to make

money or not is not really the concern, although I hope it does

However, I want you to think about the thought process behind this trade. The trade idea originated through what I consider as ‘systematic deduction’. To make such systematic deduction and find opportunities, you need to question what is happening in the market and sometimes be willing to take contrarian positions, which is exactly what we did.

‘Systematic deduction’ is one of the most popular methods market participants adapt to trade the market. However, not all systematic deductions are right, you could, of course, succumb to biases and make systematic errors while making these deductions. Nevertheless, systematic deduction is one of the other popular techniques to trade. Other popular trading techniques being –

- Trade because your gut says so
- Trade because my friend says so
- Trade because the guy on TV says so
- Trade because my broker says so

None of the above mentioned ‘approach’ to trade the market, including the ‘systematic deduction’ can really be defined as a process. These are ad-hoc methods, which cannot really be quantified or backtested.

Any approach to trade where you cannot really define ‘the approach’ as a process is not considered as a trading system.

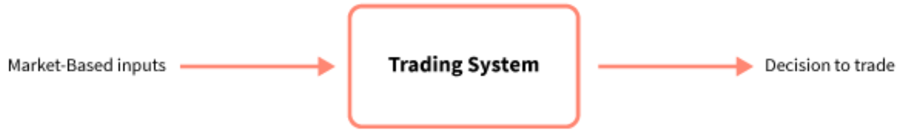
On the contrary, if you can define the approach and can quantify the process to trade the market, then you are essentially talking about a ‘Trading System’, which is exactly the focus of this module.

1.2 – Trading system – the Holy Grail?

The moment you talk about a trading system, people generally tend to think of these systems as a sure shot technique to make money, or in other words, they approach these systems as a money-making machine. They expect profits to roll from the first trade itself. Unfortunately, it does not really work that way.

Remember, a trading system receives a bunch of inputs from your end, performs a set of task, and gives you an output. Based on the output, you then decide (or the system itself decides) if this is a trade worth taking or not.

Here is how you can visualize this –



If you realize, for the trading system –

1. You give the system the inputs
2. You design the system
3. You decide to trade or not to

So the onus of making money really depends on you. The advantage of a trading system, however, is that – you only have to decide the logic once and then just follow the system that you've designed.

Of course, as you may have sensed, I've dumbed down the journey of a trading system to a large extent, and this is just to give you a perspective at this stage.

1.3 – What to expect from this module?

The trading systems that we will discuss in this module will be complete, in the sense, it will have –

1. The logic, which is the core of the trading system
2. Input parameters
3. Interpreting the output
4. The decision to trade or not

At this point, I've planned to write about the following 4 trading systems –

1. Pair trading
2. Volatility based Delta hedging
3. Calendar spreads
4. Momentum strategy (Portfolio approach)

There two techniques to pair trade – a simple approach based on correlations and a slightly complex approach using statistical concepts – both of which we will explore. Of course, as we proceed, I may try and add other trading systems as well.

However, this module will not include the 'backtest' bit. The onus is on you to backtest the system and figure out if the system works for you or not. You will have to take the rules of the system and figure out how many times in the past it has worked and if it has worked, what kind of profitability pattern the system is showcased.

Remember, no trading system is complete without having the backtesting results. The only reason why I'm not including the backtesting part is that I lack programming skills. Some of these systems can be efficiently backtested if you can manage to write a piece of code. When these systems were developed, I was fortunate enough to have a fellow trader with programming skills, hence I was in a position to get greater insights into these systems. I must also tell you that these were fairly competent systems to trade – and I presume they still are.

Of course, the market conditions have changed, hence a fresh set of backtesting is justified.

However, the broader objective of this module is to showcase different systems and give you insights into how systems are developed. Hopefully, this will inspire you to develop your own system and perhaps works out to be your own money making machine!

With this hope let us proceed – onwards to Pair trading!

PS: The short strangle on Bank Nifty worked out quite well

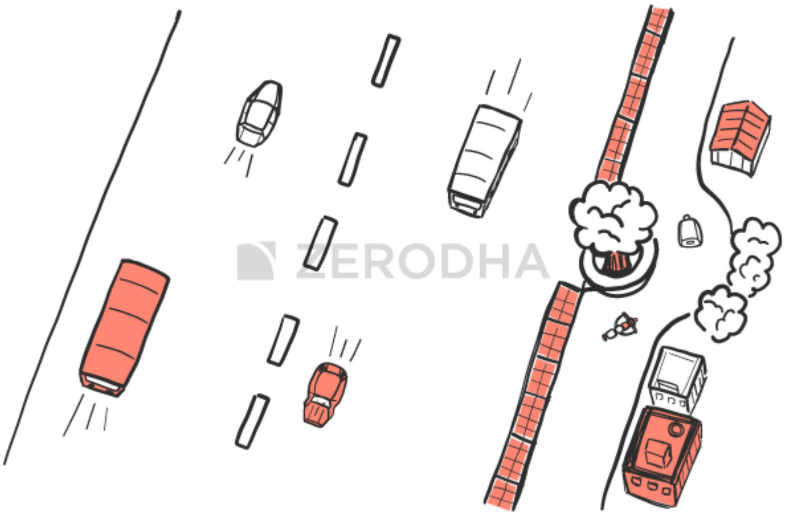
CHAPTER 2

Pair Trading logic

2.1 – The idea

If you have ever been on an interstate highway, then you would have noticed that the highway usually includes the main highway, on which the vehicles zoom by at full speed. On either side of the highway, it is common to find a single road, which is often called the service road. The service road is used to give access to private driveways, shops, houses, industries or farms. These service roads are also known as the local-express lanes. The service road and the highway usually run parallel to each other for the entire length.

Now imagine this – assume a new highway and service road is being commissioned. The road contractor has stated the work of laying down the highway and service road. At one point, on this new service road, the contractor encounters a small little tree. Now, for whatever reason, the road contractor decides not chop off the tree but instead circumvent it by taking a small deviation from the tree and get back on track to run parallel to the highway.



The road gets built this way, and people start using it. What do you make of it?

If you think about it – the two roads run parallel to each other, for the entire stretch. At any part, if the highway is inclined, so would the service road. If the highway goes down, so would the service road. If the highway crosses a river, so would the service road. So on and so forth. So for all practical purposes, the two roads 'behave' somewhat identically, except at that point where the tree briefly obstructed the path on the service road.

Let's take this a step further and break it down into variables –

1. Entities – Highway and the service road
2. Relationship – The two entities are defined by their parallelity. What happens to one entity (highway) is likely to happen to the other (service road)
3. Relationship anomaly – In an otherwise perfect world, the tree on the service road causes a brief break in the parallelity of the two roads
4. Effect of the anomaly – The anomaly is short-lived; the roads are quick to regain their relationship

I know this is a weird analogy, but if you can somehow imagine this highway, service road, and that tree, and the parallel relationship between them, then you will (hopefully) understand the underlying philosophy of pair trading.

So let me attempt to do that.

Now, just like the two roads (or entities as we defined them) i.e. the highway and service road – think about two companies which are similar, let's say – HDFC Bank and ICICI Bank.

By the way, if you pick up any classic book on Pair Trading, you will come across the example of Coca-Cola and Pepsi. Since they are not listed in India, let's go ahead with ICICI and HDFC.

1. Both these banks are very similar in every respect
2. Both are private sector banks
3. Both have similar banking products

4. Both cater to similar client base
5. Both have similar presence in the country
6. Both banks have similar regulatory constraints
7. Both banks have similar challenges in terms of running the business

So on and so forth.

Given the striking similarities between the two banks, whatever change in the business environment affects one bank, the 2nd bank should be affected in the same way. For example, if RBI increases the interest rates, then both the banks would be affected the same way and likewise when the rates are lowered.

Up to this point, we can define –

1. The entities – HDFC and ICICI
2. The relationship – similar business landscape

Given the above inference, we can make the following conclusion –

1. Because both the businesses are so alike, their stock price movement should be similar
2. On any given day, if HDFC Bank's stock price goes up, then ICICI Bank's stock price is also expected go up as well
3. If HDFC stock price comes down, then ICICI's stock price is also expected to come down

We can generalize this –

Given there is a well-established relationship between the two companies, considering all else equal, if the stock price of entity 1 moves in a certain direction, then the stock price of entity 2 is also expected to make a similar move. If not, then there could be a trading opportunity.

For example, all else equal, on a given day, ICICI stock price moves up by X% then given the relationship, HDFC is also expected to move up at least y%, but for whatever reason, assume HDFC stayed flat. Then we can go ahead and claim that ICICI stock price has moved higher than expected when compared to HDFC's stock price.

In the arbitrage world – this translates to buying the cheaper stock i.e. HDFC and selling expensive one i.e. ICICI.

In a nutshell, this is the essence of 'Pair Trading'.

Hang on a second – what about the tree on the service road and its relevance to the whole narration? Well, remember the tree caused the anomaly in an otherwise perfect 'parallel' relationship between the two roads?

Likewise, in an otherwise perfect relationship between the stock prices of two companies – an event can trigger a price anomaly – where the price of stock 1 can deviate from the price of stock 2.

An anomaly in stock prices gives us an opportunity to trade. The anomaly can happen because of anything –

1. HDFC Bank announcing quarterly results – on an immediate basis this impacts HDFC more than ICICI, hence the price relationship between the two changes, only to be realigned later
2. Likewise, with ICICI announcing its results
3. A top executive at one of these banks resigns, causing a minor dent in its stock price, while the other continues to trade regularly
4. Excessive speculation in stock 1 compared to stocks 2

Generally speaking, a price anomaly is a local event, which causes the stock price of one company reacts (or overreacts) compared to the other. I prefer to call it a local event because it affects only 1 company in our universe of two stocks J

So the relationship essentially sets the rules on how the two stock prices are related. Therefore, the bulk of the work in pair trading revolves around –

1. Identifying the relationship between two stocks

2. Quantifying their relationship
3. Tracking the behaviour of this relationship on a daily basis
4. Looking for anomalies in the price behaviour.

There are multiple ways to define these relationships between two stocks. However, the two popular techniques are based on–

1. Price spreads and ratios
2. Linear Regression

Both these techniques are different and sort of elaborate. I intend to discuss both these techniques in Varsity.

Before we close this chapter – a quick note on the history of Pair trading.

The first pair trade was executed by Morgan Stanley in the early 80's by a trader named Gerry Bamberger. Apparently, Gerry discovered the technique and kept it 'proprietary' for the longest time, until another trader called Nunzio Tartaglia, again from Morgan Stanley, popularized it.

Nunzio, at that time, had a huge following, considering he was one of the pioneers in 'Quant trading' on Wall Street. In fact, he led Morgan Stanley's prop trading desk in the 80's.

DE Shaw, the famed Hedge Fund, adopted this strategy in its initial days.

2.2 – Few closing thoughts

As you may have guessed, pair trading requires you to buy and sell two stock/assets/indices simultaneously. Many familiar with this believe that pair trading is a market neutral strategy. Market neutral, because you are both long and short at the same time. This is grossly wrong, simply because you are essentially long and short on two different stocks.

To be market neutral, you need to be – long and short, on the same underlying, at the same time. A good example here is the calendar spread. In a calendar spread, you are long and short on the same underlying expiring on two different dates.

Hence, please do not be under the impression that pair trading in market neutral. This is a trading strategy that seeks to take advantage of price differentials between two, related assets.

By simultaneously buying and selling the two assets, we are trying to profit from the "relative value" of the two securities. For this reason, I'd like to refer to Pair trading as 'Relative Value trading'.

If you think about this, in its pure sense, this is an arbitrage opportunity – we buy the undervalued security and sell the overvalued security. For this reason, some even call this the Statistical Arbitrage.

The measurement of 'undervalued' and 'overvalued' is always with respect to the one another – and the measurement technique is what we will start learning next chapter onwards.

Key takeaways from this chapter

1. The stock prices of two companies with similar business landscape tends to make similar price moves
2. The prices move can be quantified by
3. A local event (particular to 1 company) can create an anomaly in the price movement
4. When an anomaly occurs an opportunity to trade arises
5. In pair trading, you buy the undervalued security and sell the overvalued one
6. Pair trading is also called – Relative value trading or statistical arbitrage

CHAPTER 3

Pair Trading, Method 1, Chapter 1 (PTM1, C1) -Tracking Pairs

3.1 – Getting you familiar with Jargons

Like I had mentioned in the previous chapter, there are two techniques based on which you can pair trade. The first technique that we will discuss starting now, is usually referred to as the correlation based technique. I consider this as a fairly standard approach as many traders get their pair trading handholding of sorts using this approach.

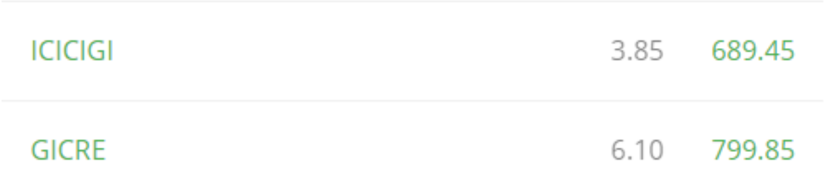
We need to learn few jargons before we get started on the actual technique, so let's get to that straight. The jargons we will talk about in this chapter are related to tracking pairs. At this stage, I just want you to know what is what. We will connect the dots as we proceed.

Spreads – The spread, is perhaps the most versatile jargon used in the trading world. For example, if you are scalping the market then the word spread refers to the Rupee differential between the bid price and the ask price. Now, if you are doing an arbitrage trade, then the word spread refers to the difference between the prices of the same asset across two different markets. In the pair trading world (actually, just correlation-based technique), the word spread refers to the difference between the closing prices of two stocks.

The spread is calculated as –

Spread = Closing value of stock 1 – closing value of stock 2

Take a look at this –



ICICIGI	3.85	689.45
GICRE	6.10	799.85

If I assume GICRE as a stock 1 and ICICIGI as stock 2, then the spread is calculated as –

Spread = 6.1 – 3.85

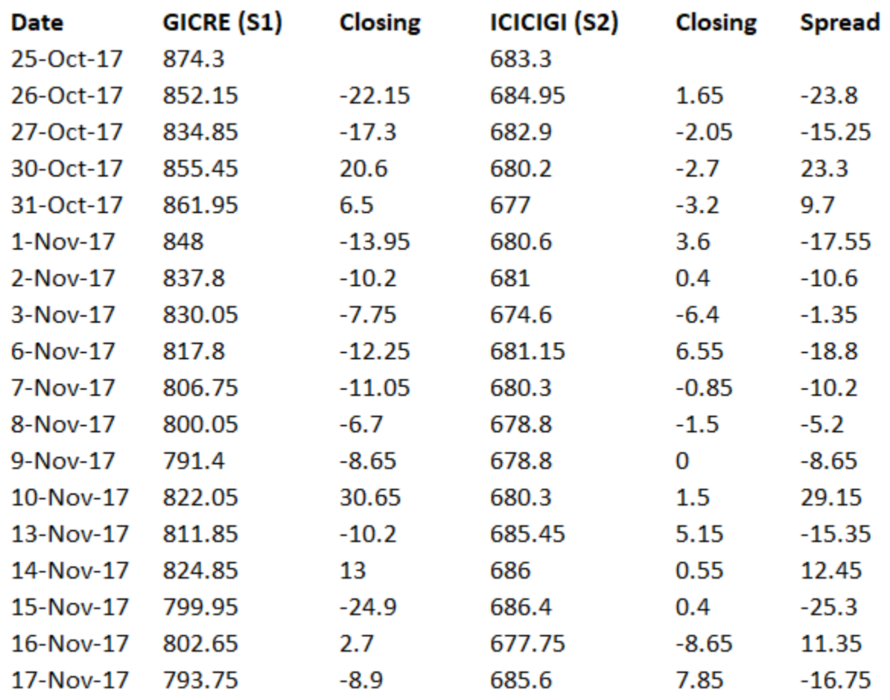
= **2.25**

Please note, both 6.1 and 3.85 represents a change in stock price with respect to the previous close. Also, both the numbers are positive here. Now, for a moment assume, the closing price of ICICIGI was negative 3.85, in this case, the spread would turn out to be –

6.1- (-3.85)

= 9.95

I've calculated the spread for the last couple of trading days, this should give you an idea of how the spread 'runs'. Also, since I've calculated the spread on a daily basis, traders refer to this as the 'historical spread'.



Date	GICRE (S1)	Closing	ICICIGI (S2)	Closing	Spread
25-Oct-17	874.3		683.3		
26-Oct-17	852.15	-22.15	684.95	1.65	-23.8
27-Oct-17	834.85	-17.3	682.9	-2.05	-15.25
30-Oct-17	855.45	20.6	680.2	-2.7	23.3
31-Oct-17	861.95	6.5	677	-3.2	9.7
1-Nov-17	848	-13.95	680.6	3.6	-17.55
2-Nov-17	837.8	-10.2	681	0.4	-10.6
3-Nov-17	830.05	-7.75	674.6	-6.4	-1.35
6-Nov-17	817.8	-12.25	681.15	6.55	-18.8
7-Nov-17	806.75	-11.05	680.3	-0.85	-10.2
8-Nov-17	800.05	-6.7	678.8	-1.5	-5.2
9-Nov-17	791.4	-8.65	678.8	0	-8.65
10-Nov-17	822.05	30.65	680.3	1.5	29.15
13-Nov-17	811.85	-10.2	685.45	5.15	-15.35
14-Nov-17	824.85	13	686	0.55	12.45
15-Nov-17	799.95	-24.9	686.4	0.4	-25.3
16-Nov-17	802.65	2.7	677.75	-8.65	11.35
17-Nov-17	793.75	-8.9	685.6	7.85	-16.75

As you can see, the spread varies on a daily basis. Also, here is an interesting (general) observation –

1. The spread expands if the closing value of S1 is positive and S2 is negative
2. The spread contracts if the closing value of S1 is positive and S2 is also positive

Of course, there are other possible combinations which lead to the expansion of contraction of the spreads. More on this later.

Differential – Unlike spreads, the differential measures the difference in the stock prices. The differential measures the absolute difference in the closing stock prices of two stock. The formula is as below –

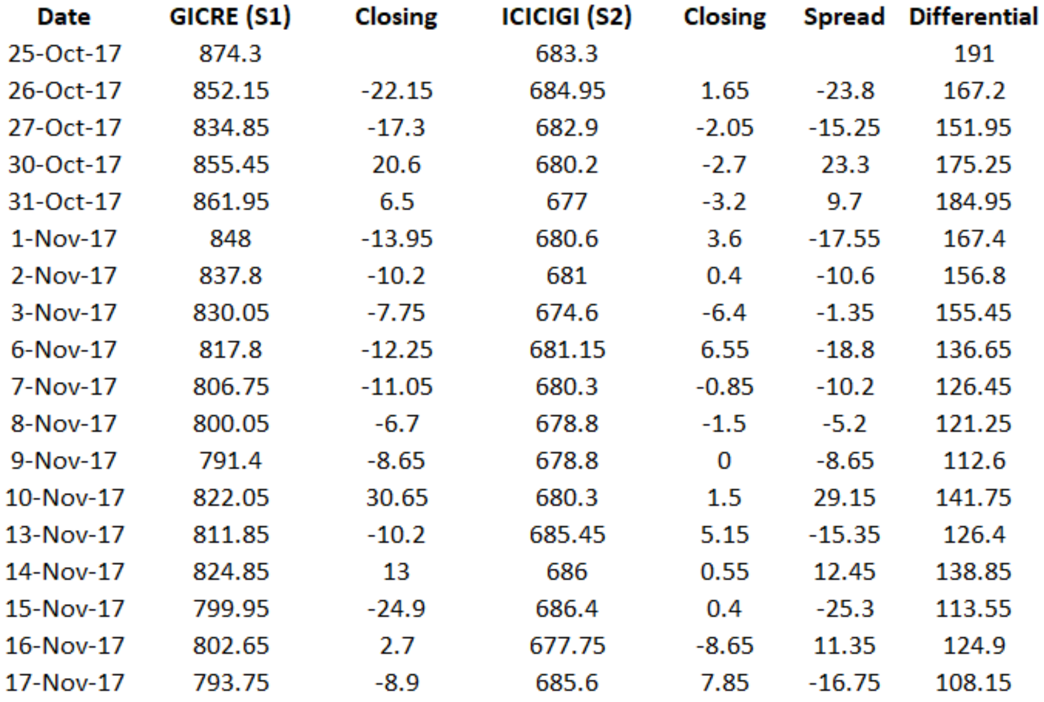
Differential = Closing Price of Stock 1 – Closing Price of Stock 2

So if a stock 1 has closed at Rs.175 and stock 2 has closed at 232, the differential is –

175 – 232

= **– 57**

As you may have guessed, you can run this as a time series and calculate this on a daily basis, I've done this for GICRE and ICICIGI –



Date	GICRE (S1)	Closing	ICICIGI (S2)	Closing	Spread	Differential
25-Oct-17	874.3		683.3			191
26-Oct-17	852.15	-22.15	684.95	1.65	-23.8	167.2
27-Oct-17	834.85	-17.3	682.9	-2.05	-15.25	151.95
30-Oct-17	855.45	20.6	680.2	-2.7	23.3	175.25
31-Oct-17	861.95	6.5	677	-3.2	9.7	184.95
1-Nov-17	848	-13.95	680.6	3.6	-17.55	167.4
2-Nov-17	837.8	-10.2	681	0.4	-10.6	156.8
3-Nov-17	830.05	-7.75	674.6	-6.4	-1.35	155.45
6-Nov-17	817.8	-12.25	681.15	6.55	-18.8	136.65
7-Nov-17	806.75	-11.05	680.3	-0.85	-10.2	126.45
8-Nov-17	800.05	-6.7	678.8	-1.5	-5.2	121.25
9-Nov-17	791.4	-8.65	678.8	0	-8.65	112.6
10-Nov-17	822.05	30.65	680.3	1.5	29.15	141.75
13-Nov-17	811.85	-10.2	685.45	5.15	-15.35	126.4
14-Nov-17	824.85	13	686	0.55	12.45	138.85
15-Nov-17	799.95	-24.9	686.4	0.4	-25.3	113.55
16-Nov-17	802.65	2.7	677.75	-8.65	11.35	124.9
17-Nov-17	793.75	-8.9	685.6	7.85	-16.75	108.15

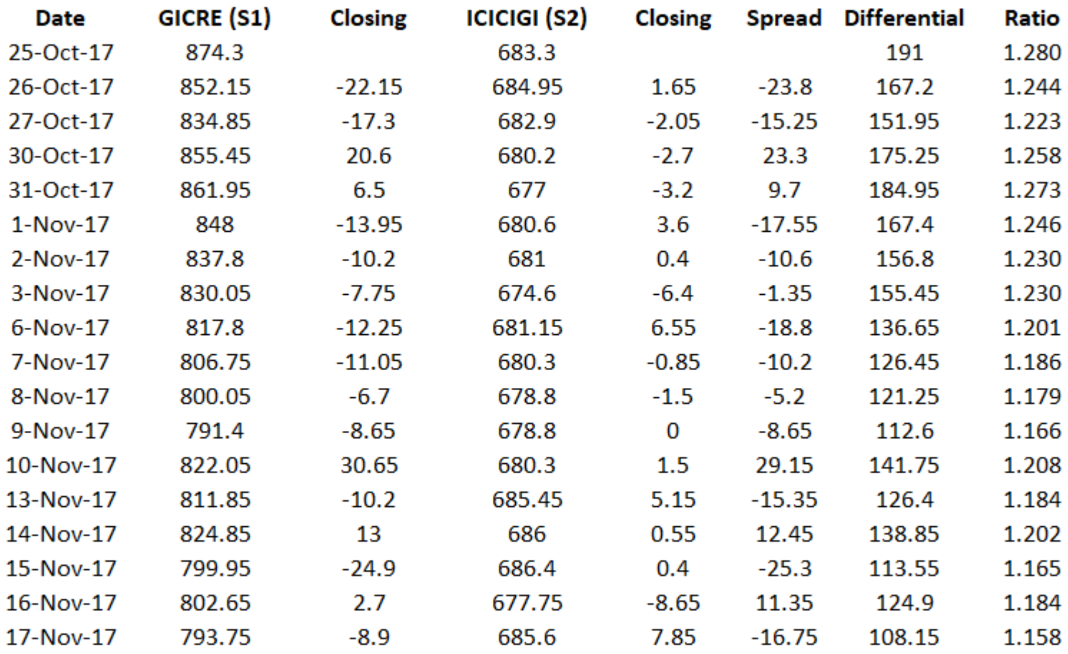
Here is something you need to know about differentials – if you are using spreads to track pairs, then you can use it on an intraday basis. But unlike spreads, the 'differentials' is not a great technique to track pairs on an intraday basis, its best used at an end of day basis.

Of course, more on these things later. For now, let's just focus on busting some jargons.

Ratio – I find the ratio bit quite interesting. The ratio is essentially dividing the stock price of stock 1 over the price of stock 2. Or it can be the other way round as well.

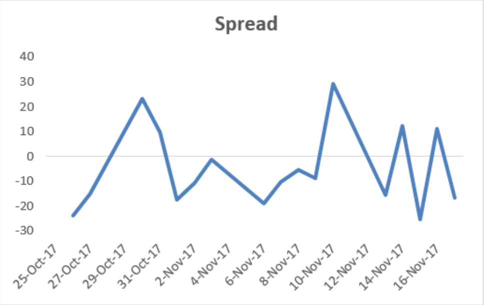
Ratio = Stock Price of stock 1 / stock price of stock 2

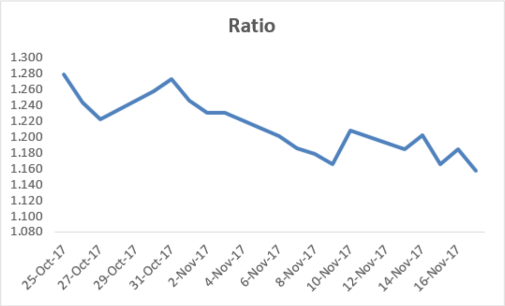
I've calculated the ratio of the same two stocks, here is how it looks –

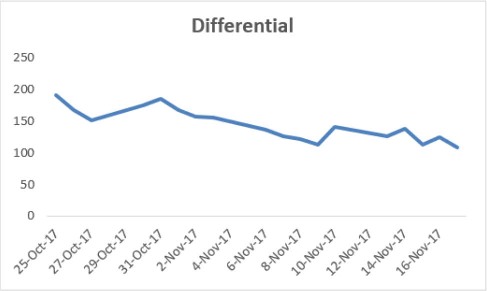


Date	GICRE (S1)	Closing	ICICIGI (S2)	Closing	Spread	Differential	Ratio
25-Oct-17	874.3		683.3			191	1.280
26-Oct-17	852.15	-22.15	684.95	1.65	-23.8	167.2	1.244
27-Oct-17	834.85	-17.3	682.9	-2.05	-15.25	151.95	1.223
30-Oct-17	855.45	20.6	680.2	-2.7	23.3	175.25	1.258
31-Oct-17	861.95	6.5	677	-3.2	9.7	184.95	1.273
1-Nov-17	848	-13.95	680.6	3.6	-17.55	167.4	1.246
2-Nov-17	837.8	-10.2	681	0.4	-10.6	156.8	1.230
3-Nov-17	830.05	-7.75	674.6	-6.4	-1.35	155.45	1.230
6-Nov-17	817.8	-12.25	681.15	6.55	-18.8	136.65	1.201
7-Nov-17	806.75	-11.05	680.3	-0.85	-10.2	126.45	1.186
8-Nov-17	800.05	-6.7	678.8	-1.5	-5.2	121.25	1.179
9-Nov-17	791.4	-8.65	678.8	0	-8.65	112.6	1.166
10-Nov-17	822.05	30.65	680.3	1.5	29.15	141.75	1.208
13-Nov-17	811.85	-10.2	685.45	5.15	-15.35	126.4	1.184
14-Nov-17	824.85	13	686	0.55	12.45	138.85	1.202
15-Nov-17	799.95	-24.9	686.4	0.4	-25.3	113.55	1.165
16-Nov-17	802.65	2.7	677.75	-8.65	11.35	124.9	1.184
17-Nov-17	793.75	-8.9	685.6	7.85	-16.75	108.15	1.158

The Ratio as you can see is a bit more consistent (or at least appears) when calculated as a time series. I've represented all the three variables on graph –







So what are these things that we just looked at – spread, differential, and ratios and how are they related to pair trading?

Well, as you can imagine, these are the different variable which helps us measure or quantify the relationship between two stocks, which we consider as pairs. The graph tells us how the two stocks move with respect to each other. For instance, if we consider the spread, we know it expands if the closing value of S1 is positive and S2 is negative and the spread contracts if the closing value of S1 is positive and S2 is also positive.

Likewise, in the ratio – the ratio between two stocks decrease if the stock prices of both the stock decline and the ratio increases if the stock prices of both the stocks increases. Of course, there are other variations possible – for example, the ratio can increase if stock 1 declines heavily and stock 2 stays flat or the other way round. Alternatively, stock 2 can increase a lot more compared to stock 1 or the other way round J

Confusing isn't it?

Hence, for this reason, we need to look at the chart of the variable we are following, the variable could be spread, differential, or the ratio. We need to track the movement of the variable and figure out if the spread is expanding or contracting. This leads us to the next two jargons.

Divergence – If the ratio or the spread between the two stocks is expected to move apart or alternatively, you expect the graph to move up, then this translates to something called a divergence. When you expect your variable to diverge, you can make money (or at least attempt to make) by setting up a divergence trade.

Convergence – If the ratio or the spread between the two stocks is expected to move closer or alternatively, you expect the graph to move down, then this translates to something called as a convergence. When you expect your variable to converge, you can make money (or at least attempt to make) by setting up a convergence trade.

Now here is the big question – what makes you believe the variable can either converge or diverge? When do you decide to set up a trade? What are the triggers? How do you set up a trade? What if the trade does not work out? What is the stop-loss for such trades?

Well, even before we answer these questions, how do we qualify two stocks as a pair? Just because two stocks belong to the same sector, does that mean they qualify as a pair? For instance, does ICICI Bank and HDFC Bank qualify as a pair because they both belong to private sector banking?

To qualify two stocks as a pair we need to rely upon the good old statistical measure, called the 'Correlation'. I guess, we have discussed correlation multiple times on varsity. Here is a quick explanation –

Correlation between two variables gives us a sense of how two variables move with respect to each other. Correlation is measured as a number which varies between -1 to +1. For example, if the correlation between two stocks is +0.75, then it tells us two things –

1. The plus preceding the number tells us that they both are positively correlated i.e. when they move in the same direction
2. The actual number gives us a sense of the strength of this movement. In a loose sense, the closer it is to +1 (or -1) the higher is the tendency for the two variable to move in tandem.
3. A correlation of 0 suggests that the two variables are not related to each other.

From the above, we know a correlation of +0.75 suggests that the two variables move not only in the same direction but also tend to move together closely. Note, the correlation does not suggest the extent of the move, all it suggests is that the move in the same direction is likely to happen. For example, if Stock A moves 3%, and the correlation between stock A and stock B is +0.75, then it does not mean that Stock B will also move by 3%, all that the correlation suggests is that Stock B will move up positively, just like Stock A.

But, there is another twist here – suppose stock A and Stock B are correlated at 0.75, and the daily average return on Stock A and Stock B is 0.9% a 1.2%, then it can be said that on any given day, if Stock A moves above its daily average return of 0.9%, then stock B is also likely to move higher than its daily average return of 1.2%.

Likewise, a correlation of -0.75 indicates that the two variables move in opposite direction (-ve sign) but they both tend to move in opposite direction. Suppose stock A moves up by +2.5%, then by virtue of correlation we know that Stock B is likely to come down, but by what degree will it come down will not be known.

While we are at it, one more point on correlation. This bit is only for those interested in the math part of correlation. The correlation data makes sense only if the data series is 'stationary around the mean'. What does this mean? – Well, it simply means that the data set should be sticking close the average values.

Keep this line 'stationary around the mean' in the back of your mind, don't forget it. This will come back to again, when we discuss the 2nd technique to pair trade, much later in this module.

We will proceed with correlation as a measure to understand how tightly two stocks are coupled. In the next chapter, we will figure out how to calculate two different varieties of correlations.

For now, I want you to be clear on Spread, Differentials, Ratios, Divergence Trading, Convergence Trading, and Correlations!

Download the Excel sheet used in this chapter **here**.

Key takeaways from this chapter

1. Spread measures the difference between the closing values of two stocks
2. Differentials measures the difference between the closing prices of two stock
3. Ratio between the two stocks essentially requires you to divide stock 1 over stock 2
4. Divergence is when you expect the two stocks to move apart
5. Convergence is when you expect the two stocks closer to each other
6. Correlation is like a glue which tells how tightly two stocks move together.

CHAPTER 4

PTM1, C2 – Pair stats

4.1 – Correlation and its types

I have to mention this at this point. The pair trading technique we are discussing now is discussed in a book called, ‘Trading Pairs’, by Mark Whistler. I like this book for the fact that it got me hooked to Pair trading and over time as my interest grew, I explored the strategy beyond Mark Whistler’s techniques. Needless to say, I will discuss those techniques later in this module. At this point, my intention is to take you through the exact learning path I underwent learning pair trading.

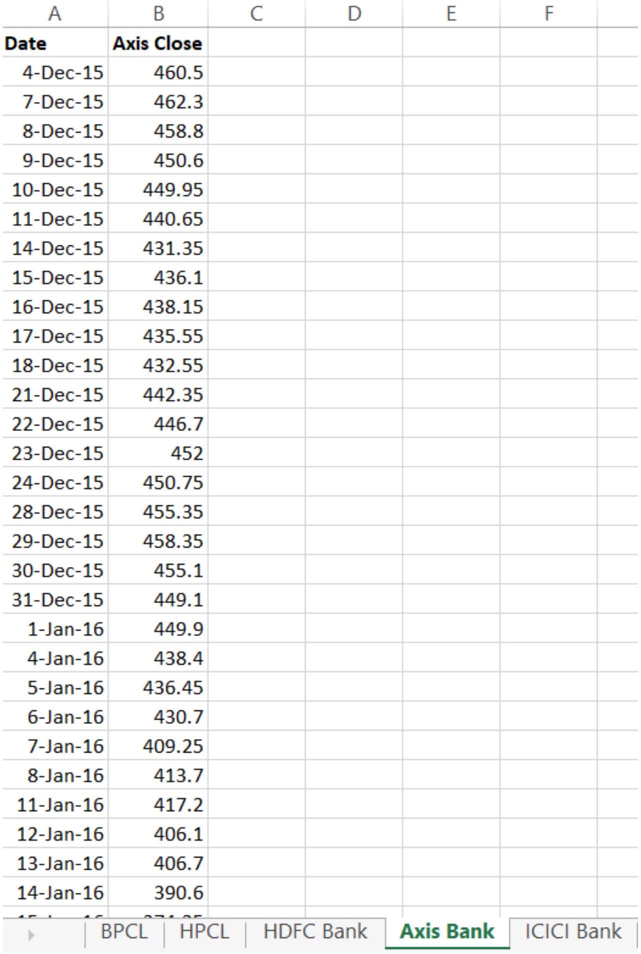
Towards the end of the previous chapter, we introduced the concept of correlation and the way one can analyse the correlation values. We will take that discussion forward now and understand how to calculate the correlation between two stocks, on excel. As you may have guessed by now, the calculation of Correlation between two stocks is the key in pair trading.

For the sake of this example, I've considered Axis Bank and ICICI Bank. Both are Private sector banks and have similar business backgrounds, hence intuition says that the two stocks should be highly correlated.

At this point, I have downloaded the closing price of Axis Bank and ICICI Bank from 4th Dec 2015 to 4th Dec 2017, roughly 2 years of trading data or about 496 data points.

Before we proceed, a quick note on data –

1. Make sure you are dealing with the same number of data points. For example, if you have 400 data point for Stock A, then you need to ensure you have the same number of data points for Stock B, corresponding to same dates.
2. Make sure the data is cleaned for corporate actions such as bonus/splits etc.



A	B	C	D	E	F
Date	**Axis Close**				
4-Dec-15	460.5				
7-Dec-15	462.3				
8-Dec-15	458.8				
9-Dec-15	450.6				
10-Dec-15	449.95				
11-Dec-15	440.65				
14-Dec-15	431.35				
15-Dec-15	436.1				
16-Dec-15	438.15				
17-Dec-15	435.55				
18-Dec-15	432.55				
21-Dec-15	442.35				
22-Dec-15	446.7				
23-Dec-15	452				
24-Dec-15	450.75				
28-Dec-15	455.35				
29-Dec-15	458.35				
30-Dec-15	455.1				
31-Dec-15	449.1				
1-Jan-16	449.9				
4-Jan-16	438.4				
5-Jan-16	436.45				
6-Jan-16	430.7				
7-Jan-16	409.25				
8-Jan-16	413.7				
11-Jan-16	417.2				
12-Jan-16	406.1				
13-Jan-16	406.7				
14-Jan-16	390.6				

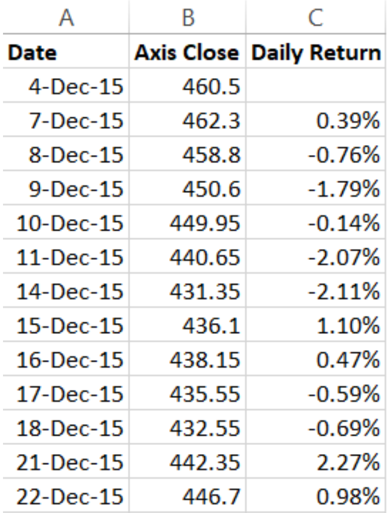
As you can see from the above image, besides ICICI and Axis, I have also downloaded the data for BPCL, HPCL, and HDFC Bank. You can use this data to build and test other correlations.

Anyway, at this stage, the only data we have is the date and the closing price of the stock. We will go ahead and calculate the daily returns. I guess you are familiar with the daily return calculation; we have discussed this several time in the previous module.

The daily return can be calculated as

= [today's closing price / previous day's closing price] – 1

I've calculated this for both ICICI and Axis Bank –



A	B	C
Date	**Axis Close**	**Daily Return**
4-Dec-15	460.5	
7-Dec-15	462.3	0.39%
8-Dec-15	458.8	-0.76%
9-Dec-15	450.6	-1.79%
10-Dec-15	449.95	-0.14%
11-Dec-15	440.65	-2.07%
14-Dec-15	431.35	-2.11%
15-Dec-15	436.1	1.10%
16-Dec-15	438.15	0.47%
17-Dec-15	435.55	-0.59%
18-Dec-15	432.55	-0.69%
21-Dec-15	442.35	2.27%
22-Dec-15	446.7	0.98%

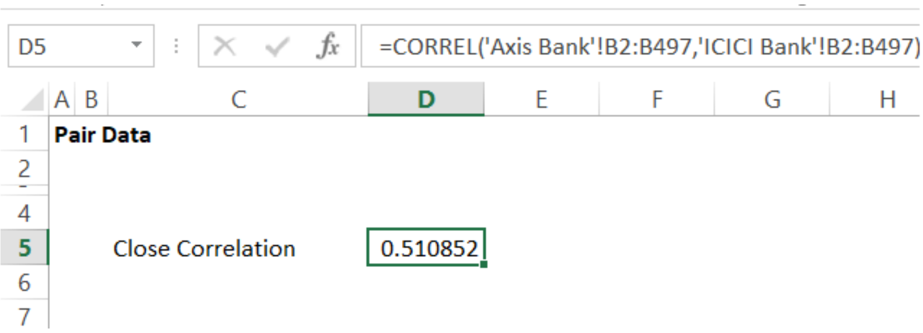
Now, correlation can be calculated on the basis of two parameters –

1. The daily closing price
2. The daily return series

The daily closing price correlation requires you to calculate the correlation based on the closing prices of two stock. I'm not a big fan of calculating correlation on closing prices, but then let's just go ahead and do this for time being.

To do this in excel, simply use the '=Correl()', function on the daily closing prices. I'm running this calculation on a new sheet, which is labelled it as 'Pair Data'.

Here is the snapshot –


D5

=CORREL('Axis Bank'!B2:B497,'ICICI Bank'!B2:B497)

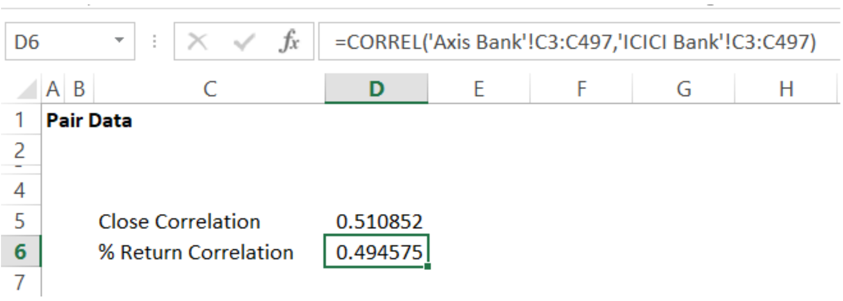
Pair Data

Close Correlation 0.510852

The correlation between the closing prices of ICICI Bank and Axis bank is 0.51. Not particularly a great correlation, but we can live with this for now. Do recollect, our gut said the two banks could be highly correlated as they have similar business backgrounds, but

the number is painting a slightly different picture

We will now run the correlation on the daily % return series for the two stock. I've already calculated this % return, I'll just have to run the correl function now.


D6

=CORREL('Axis Bank'!C3:C497,'ICICI Bank'!C3:C497)

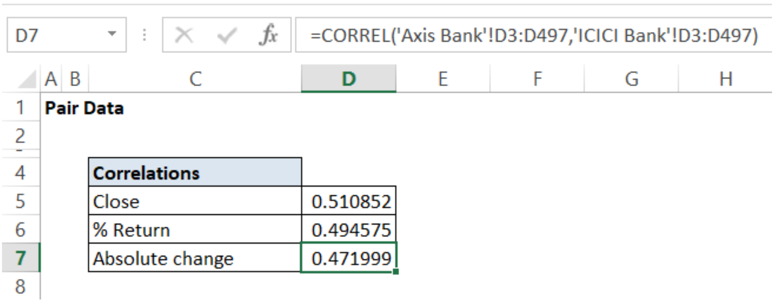
Pair Data

Close Correlation 0.510852

% Return Correlation 0.494575

Again, not a very encouraging number, but that is ok for now.

Some traders, run the correlation on the absolute per day change calculated as 'Today's stock price – yesterday's stock price'. Again, I'm not a big fan of this. But let me just go ahead and introduce the same to you –



=CORREL('Axis Bank'!D3:D497,'ICICI Bank'!D3:D497)

Pair Data

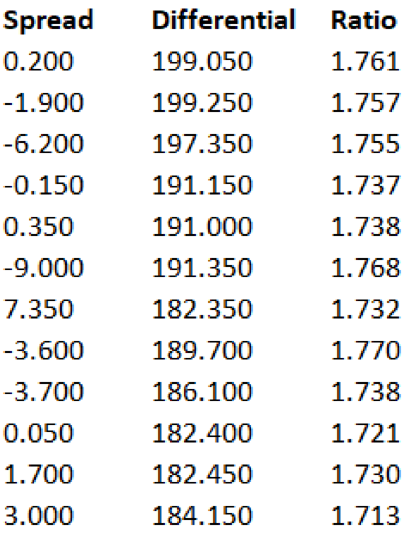
Correlations	
Close	0.510852
% Return	0.494575
Absolute change	0.471999

In all the above calculations, I've run the correlation of Axis Bank versus ICICI Bank, the results obtained will be same if I had opted to calculate the correlation of ICIC Bank versus Axis. Generally speaking, the correlation between A and B is the same as Correlation between B and A.

In this method of trading pairs, the correlation number is considered sacred. Ideally speaking, the number should be above 0.75. Clearly, that is not the case with ICICI and Axis, but then as I mentioned earlier, we can live with it.

4.2 – Setting up the datasheet

In the previous chapter, we discussed three variables concerning the pairs namely the spread, differential, and the ratios. Let us go ahead and calculate these variables on the two stocks we are studying. We will do this on a separate sheet within the same workbook and name the sheet as the 'Data Sheet'. Here is the snapshot –



Spread	Differential	Ratio
0.200	199.050	1.761
-1.900	199.250	1.757
-6.200	197.350	1.755
-0.150	191.150	1.737
0.350	191.000	1.738
-9.000	191.350	1.768
7.350	182.350	1.732
-3.600	189.700	1.770
-3.700	186.100	1.738
0.050	182.400	1.721
1.700	182.450	1.730
3.000	184.150	1.713

The calculation of these variables is quite straightforward, I've explained this in the previous chapter.

Different types of Pair Trading works at different complexities levels. We will deal with basic stats for this version of pair trading. Given this, we will now define 3 most commonly used statistic variables.

4.3 – Basic stats

I'll discuss 3 basic statistical terms at this stage. These are basic terms which play a very crucial role in pair trading. I'm fairly certain that you'd have learned these in your high school math, even otherwise this is quite basic and you can pick it up anytime.

To help you understand these jargons better, I've come up with a set of arbitrary runs scored by batsmen across 10 cricket matches –

Match	Runs scored
1	72
2	65
3	44
4	100
5	82

6	55
7	100
8	23
9	51
10	34

Mean – Also called the arithmetic average, represents the average of a set of numbers. You can calculate the average by taking the sum of all the observations by the total number of observations.

So if I were to find the average in the above example, I'd total up all the scores and divide it by 10 (10 being the total number of observations).

Mean (Average) = 626/10

=62.6

On excel, you can simply use the '=Average ()' function to calculate the average of any set of numbers.

Median – The median number represents the middle number of the data series when the data series is arranged in its numerical order. If there are even set of numbers (which is the case here), then we have to take the average of the middle two numbers to calculate the mean. However, if there are an odd number of data points, then we simply take the middle data point as the median.

So let me rearrange the data points in its numerical order –

23, 34, 44, 51, 55, 65, 72, 82, 100, 100

Since there are even numbers of observation, I'll take the middle two numbers i.e. 55 and 65, their average represents the median.

Median = (55 + 65)/2

=**60**

The excel function to calculate median is '=Median()'.

The mean and median when viewed together gives a sense of the trend. More on this later.

Mode – The mode of a data series is simply that data point which occurs the most number of times in the series. Clearly, 100 is repeating twice, with no other number appearing more than once, and that makes it the mode of the data series.

The excel function to calculate Mode is '=Mode()'.

In the next chapter, we will use these function in excel and understand its relevance to pair trading.

Download the excel sheet used in this chapter **here**.

Stay tuned.

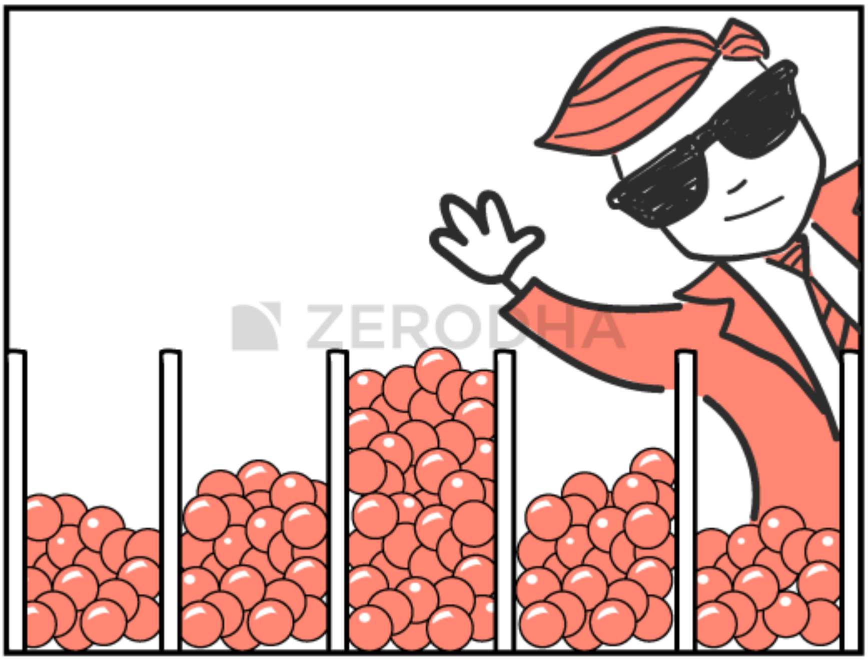
Key takeaways from this chapter

1. Care has to be taken to ensure the data is clean and adjusted for corporate actions
2. Close correlation is the correlation when calculated on the closing prices of stocks
3. The % return correlation is the correlation when calculated on the daily returns of the stock
4. Mean is the arithmetic average of the data series
5. Median is the middle observation of a data series.

6. If the data series has even number of observations, then the median is the average of the middle two observations
7. If the data series has odd number of observations, then the median is the middle observation
8. The mode of a data series is that value which repeats the highest number of times
9. The mean and median, when viewed together to each other, offers great insight into the data trend.

CHAPTER 5

PTM1, C3 – Pre trade setup



5.1 – Revisiting the Normal Distribution

If you have been a regular reader on Varsity, then chances are you'd have come across the discussion on Normal Distribution in the Options Module. If you're not, then I'd strongly suggest you read up this chapter on **Normal distribution**.

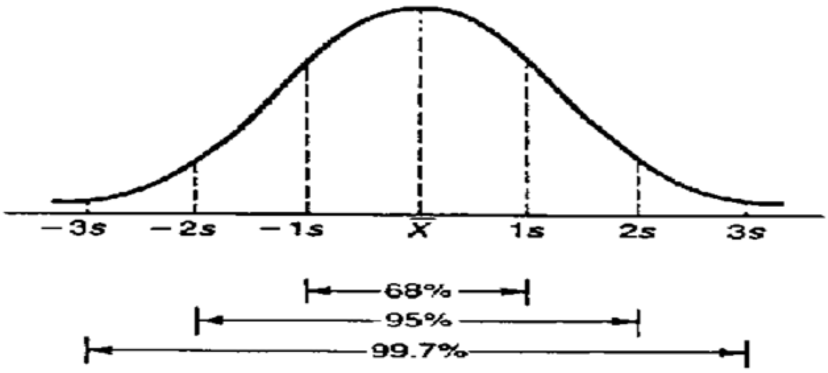
This is a very important topic, I'd suggest you spend some time reading about it before you proceed. We will use the concept of Normal Distribution in both the techniques of Pair Trading, i.e. the Mark Whistler's Pair Trading technique, and the other technique we will discuss later on in this module. Given the central role it plays, you should spend time reading about it.

I'm reproducing the central theme around Normal distribution, this should serve as a quick refresher for people who are familiar with Normal Distribution, but for those who are not, I hope this does not demotivate you from reading the chapter on Normal distribution –

The general theory around the normal distribution which you should know –

- Within the 1st standard deviation, one can observe 68% of the data
- Within the 2nd standard deviation, one can observe 95% of the data
- Within the 3rd standard deviation, one can observe 99.7% of the data

The following image should help you visualize the above –



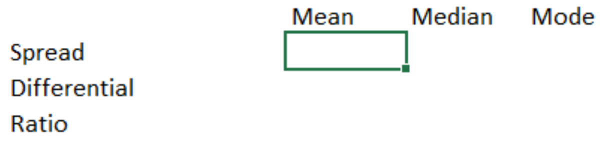
Of course, there are other forms in which the data gets distributed – distribution such as uniform, binomial, exponential distribution etc. This is just for your information.

5.2 – Descriptive Statistics

In the previous chapter, we discussed three basic statistical metrics namely the Mean, Median, and Mode. We will now calculate these metrics on the pair data i.e. the differential, spread, and ratio which we computed in the previous chapter. We will do these calculations using the excel functions.

Please note, I'm continuing on the excel that we were working on in the previous chapter, needless to say, you can download the updated excel from the link provided towards the end of the chapter.

The sheet is set up as below –

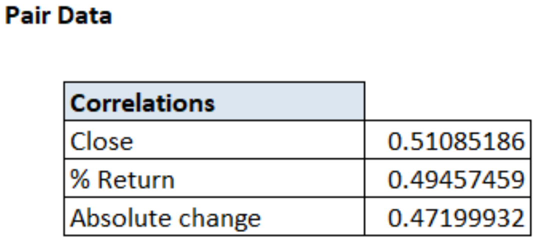


	Mean	Median	Mode
Spread			
Differential			
Ratio			

The Excel functions are as follows –

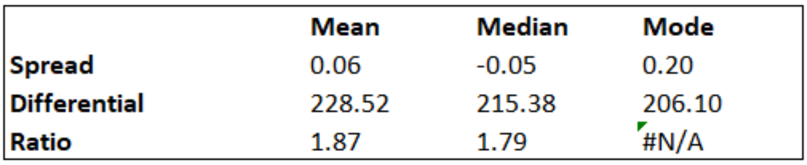
1. Mean – ‘=average()’
2. Median – ‘=median()’
3. Mode – ‘=mode.mult()’

And the numbers are as below –



Pair Data

Correlations	
Close	0.51085186
% Return	0.49457459
Absolute change	0.47199932



	Mean	Median	Mode
Spread	0.06	-0.05	0.20
Differential	228.52	215.38	206.10
Ratio	1.87	1.79	#N/A

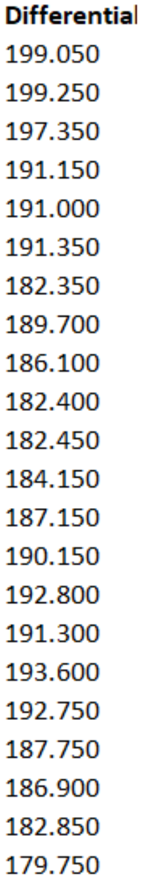
As you may notice, the correlation numbers were calculated in the previous chapter.

We now have the data setup. We need to add one key variable here and that would be the standard deviation. Again, standard deviation as a concept has been explained in Varsity earlier. I’d suggest you **read this chapter** to understand Standard Deviation better. Here is the summary though –

Standard Deviation simply generalizes and represents the deviation from the average. Here is the textbook definition of SD “*In statistics, the* ***standard deviation*** *(SD, also* *represented by the Greek letter sigma, ⌧) is a measure that is used to quantify the amount of variation or dispersion of a set of data values*”.

So in a sense, Standard Deviation gives us a sense of variability of the data or in other words, help us understand how widely the data set is spread out. Let me try and put this in the context of the Pair data we are dealing with.

The differential data which we computed a while ago is something like this –



Differential
199.050
199.250
197.350
191.150
191.000
191.350
182.350
189.700
186.100
182.400
182.450
184.150
187.150
190.150
192.800
191.300
193.600
192.750
187.750
186.900
182.850
179.750

Together there are 496 differential data points and earlier in this chapter, we have even calculated the average value across these data points i.e. 228.52.

Now, what if I were to ask you to help me understand the variability of these data points from its average value? Or a better question to ask – why would I need to know the variability of the data points from its average value?

Well, if we don't know the variability of the data, then there is no way we can make an intelligent assessment of the behaviour of the data set. For example, when the 498th data is generated, we will know if this value is around the mean or within the range it varies.

This, in fact, forms the crux of pair trading.

Standard Deviation helps us measure this variation.

While I personally think standard deviation is good enough, there are traders who would also like to calculate another variable called the 'Absolute Deviation'. Both standard deviation and absolute deviation help us understand the variability of the data. But they differ in terms of the way do they data is treated.

I was looking at the explanation to help you understand the difference between standard deviation and absolute deviation, and I found the following on Investopedia, which I think is quite nice. I'm taking the liberty of reproducing the content here –

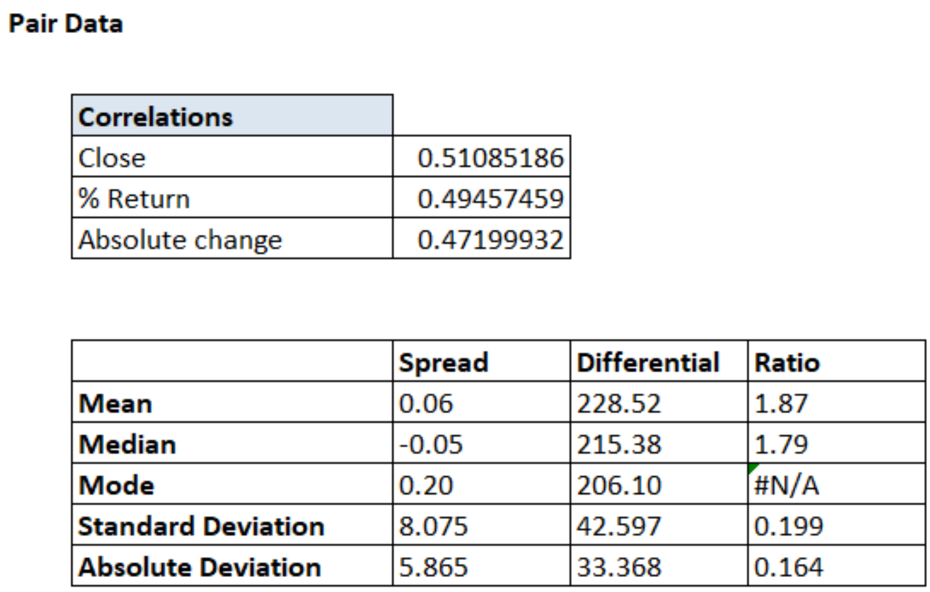
"While there are many different ways to measure variability within a set of data, two of the most popular are standard deviation and average deviation. Though very similar, the calculation and interpretation of these two differ in some key ways. Determining range and volatility is especially important in the finance industry, so professionals in areas such as accounting, investing and economics should be very familiar with both concepts.

Standard deviation is the most common measure of variability and is frequently used to determine the volatility of stock markets or other investments. To calculate the standard deviation, you must first determine the variance. This is done by subtracting the mean from each data point and then squaring, summing and averaging the differences. Variance in itself is an excellent measure of variability and range, as a larger variance reflects a greater spread in the underlying data. The standard deviation is simply the square root of the variance. Squaring the differences between each point and the mean avoids the issue of negative differences for values below the mean, but it means the variance is no longer in the same unit of measure as the original data. Taking the root of the variance means the standard deviation returns to the original unit of measure and is easier to interpret and utilize in further calculations.

The average deviation, also called the mean absolute deviation, is another measure of variability. However, average deviation utilizes absolute values instead of squares to circumvent the issue of negative differences between data and the mean. To calculate the average deviation, simply subtract the mean from each value, then sum and average the absolute values of the differences. The mean absolute value is used less frequently because the use of absolute values makes further calculations more complicated and unwieldy than using the simple standard deviation."

We will go ahead and compute both "Standard Deviation", and "Absolute Deviation" for all the three pair data variables.

By the way, I'm interchanging the Y-axis to Mean, Median, and Mode. The X-axis to Differential, Ratio, and Spread. Given this, the snapshots posted above will be slightly different from the one posted below, hope you won't mind my clumsy data handling skills J



Pair Data

Correlations	
Close	0.51085186
% Return	0.49457459
Absolute change	0.47199932

	Spread	Differential	Ratio
Mean	0.06	228.52	1.87
Median	-0.05	215.38	1.79
Mode	0.20	206.10	#N/A
Standard Deviation	8.075	42.597	0.199
Absolute Deviation	5.865	33.368	0.164

The excel function to calculate these variables are –

Standard Deviation – '=Stdev.p()'

Absolute Deviation – '=avedev()'

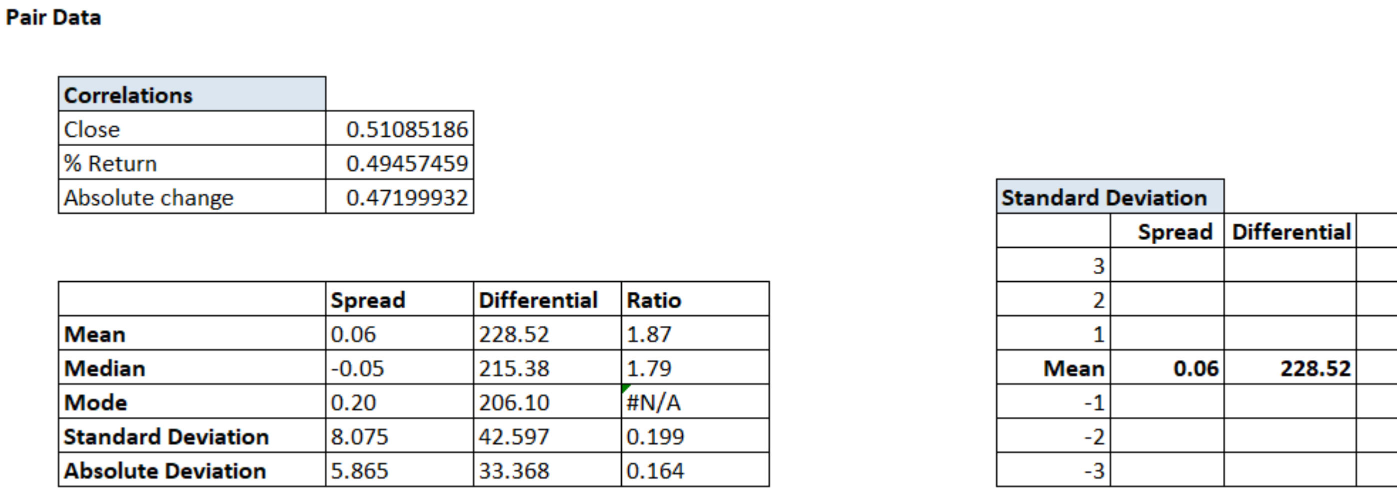
The Mean, Median, Mode, Standard Deviation, and Absolute Deviation is also known as the basic descriptive statistics.

5.3 – The Standard deviation table

The standard deviation as you know helps us get a sense of the variation in the data. We will now take this a step further and try and quantify the variation. Why do we need to do this, you may ask? Well, this will help us understand the extent of the variation from the mean value. For example, the 498th differential data could be 275, we will exactly know if 275 is way above the mean or way too below the mean.

With this information, we can choose to either buy the pair or short the pair. Of course, we will get into these details later on. For now, let us focus on quantifying the extent of the variation. In order to quantify the data point, we need to build something called as a standard deviation table.

The structure of the table is as below –



Pair Data

Correlations	
Close	0.51085186
% Return	0.49457459
Absolute change	0.47199932

	Spread	Differential	Ratio
Mean	0.06	228.52	1.87
Median	-0.05	215.38	1.79
Mode	0.20	206.10	#N/A
Standard Deviation	8.075	42.597	0.199
Absolute Deviation	5.865	33.368	0.164

Standard Deviation		
	Spread	Differential
3		
2		
1		
Mean	**0.06**	**228.52**
-1		
-2		
-3		

As you may have guessed, we are now going to calculate the values of 1, 2, and 3 standard deviations above the mean and below the mean, across spread, differential, and the ratio.

For example, let us just focus on the Spread data for now. The mean of the spread is 0.06. We also know the standard deviation (SD) is 8.075.

Therefore, the 1st SD above the mean would be –

0.064 + 8.075 = **8.139**

2nd SD –

0.064 + (2*8.075) = **16.123**

3rd SD –

0.064 + (3*8.075) = **24.288**

These are all values above the mean. We can do the same to identify the values below the mean –

-1 SD –

0.064 – 8.075 = **-8.011**

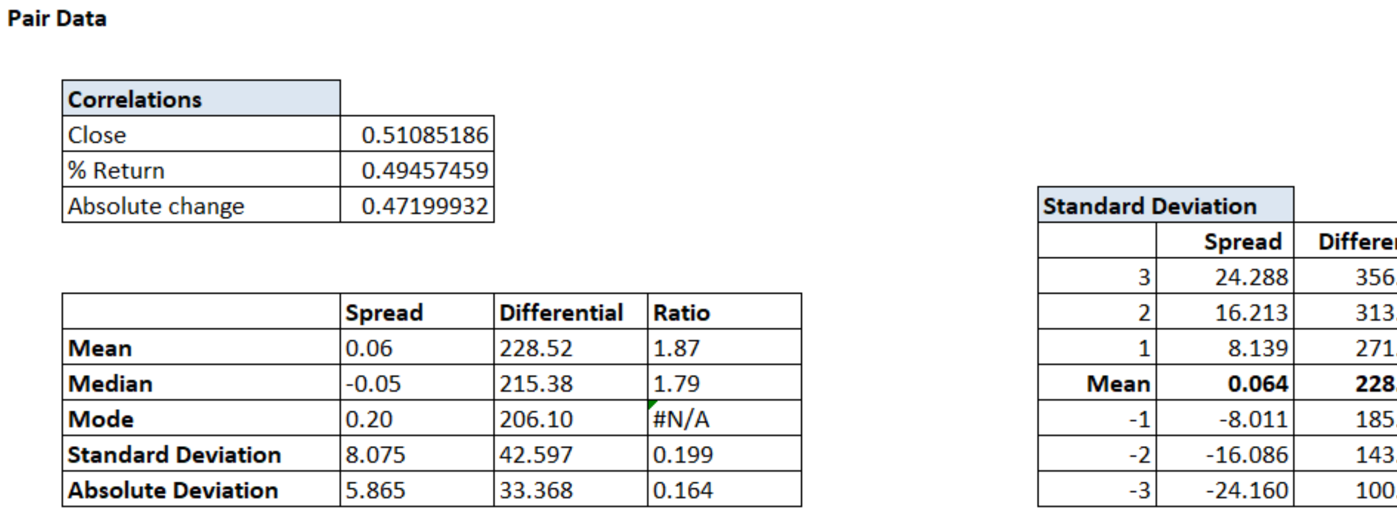
-2 SD –

0.064 – (2*8.075) = **-16.086**

-3 SD –

0.064 – (3*8.075) = **-24.160**

I've done the same math across Differential and Ratio. Here is how the table looks –



Pair Data

Correlations	
Close	0.51085186
% Return	0.49457459
Absolute change	0.47199932

	Spread	Differential	Ratio
Mean	0.06	228.52	1.87
Median	-0.05	215.38	1.79
Mode	0.20	206.10	#N/A
Standard Deviation	8.075	42.597	0.199
Absolute Deviation	5.865	33.368	0.164

Standard Deviation		
	Spread	**Differe**
3	24.288	356
2	16.213	313
1	8.139	271
Mean	**0.064**	**228**
-1	-8.011	185
-2	-16.086	143
-3	-24.160	100

So if the 498th differential data read 315, then we can quickly understand that the value is around the +2 standard deviations and with 95% confidence you could conclude that there is only 5% chance for the next set of data points to go higher than 315.

Anyway, at this stage, we have almost all the data that we need to make the assessment of the pair and probably identify if there is an opportunity to trade. In the next chapter, we will go ahead and do this. In fact, I'll start the next chapter with a quick recap of everything we have discussed so far, this is just to ensure we are all on the same page.

You can download the excel sheet used in this chapter **here**.

Signing of this chapter by wishing you all a very happy Xmas and a happy new year! Hope 2018 brings in wisdom, wealth, and peace your way.

Key takeaways from this chapter

1. Normal distribution plays a pivotal role in pair trading
2. Within the 1st standard deviation, one can observe 68% of the data
3. Within the 2nd standard deviation, one can observe 95% of the data
4. Within the 3rd standard deviation, one can observe 99.7% of the data
5. Standard deviation and absolute deviation measures the variability of the data
6. The standard deviation table gives us a sense of how the current data stands with respect to its expected variation
7. The cues to trade the pair either long or short comes from the standard deviation table.

CHAPTER 6

PTM1, C4 – The Density Curve

6.1 – A quick recap

I think a quick recap is justified at this stage, this is to ensure we are all on the same page. I'd strongly recommend you read through the recap, to ensure we are on track. I'll keep this as a pointwise recap to ensure we don't digress.

- Two companies are comparable if they have similar business background
- Business background includes factors which influence the day to day running of the business
- If two companies have similar business backgrounds, then it is reasonably safe to assume that their share prices move somewhat similarly on a day to day basis
- If the daily stock price of two comparable companies move together (and therefore their daily returns), then they do tend to have a tight correlation
- There are times when a local event can change the course of the movement in the stock price of one of the two companies, creating a pair trading opportunity
- The relationship between the stock prices of the two companies can be estimated by any of the three variables – spread, differential, or ratio
- The variables are expected to be normally distributed, hence we calculate the standard deviation of these variables, along with the basic descriptive statistics such as the mean, median, and mode.
- As a ready reckoner, we also have the standard deviation (SD) table, extending up to the 3rd SD, either sides
- Lastly, do remember we are in the process of discussing two variants of pair trading, starting with Paul Whistler's technique of Pair Trading. After this, I will discuss a slightly more complicated version of Pair Trading

So this brings us to where we are at this stage. In this chapter, we will go ahead and discuss the density curve and the eventual trigger to pair trade.

6.2 – Selecting the variable

We have come to a stage where we need to stick to one of the variables amongst Spread, Differential, and Ratio. Why just and why not all, you may ask?

Well, this is to ensure that we are sticking to a regime and not really getting confused with conflicting signals. The reason I've introduced all three variables is to showcase that there are three different possibilities. It is up to you as a trader to choose the variable that you are most comfortable with. For example, I personally prefer the ratio over the differential or spread. This is because the ratio kind of captures the market valuation of the stocks since it considers the latest stock price. Besides the ratio also gives us a quick sense of how much of Stock 1 should be bought or sold with respect to stock 2.

For example, if the price of Stock 1 is 190 and Stock 2 is 80, then the ratio of stock 1 over 2 is –

190/80

= 2.375

This implies for every 1 share of Stock 1, 2.375 shares of Stock 2 have to be transacted. We will get to the finer details later, but for now, hope you get the drift.

You are of course, free to choose any of the variable – spread, differential, or ratio. However, for the sake of this discussion, I will go ahead with the ratio.

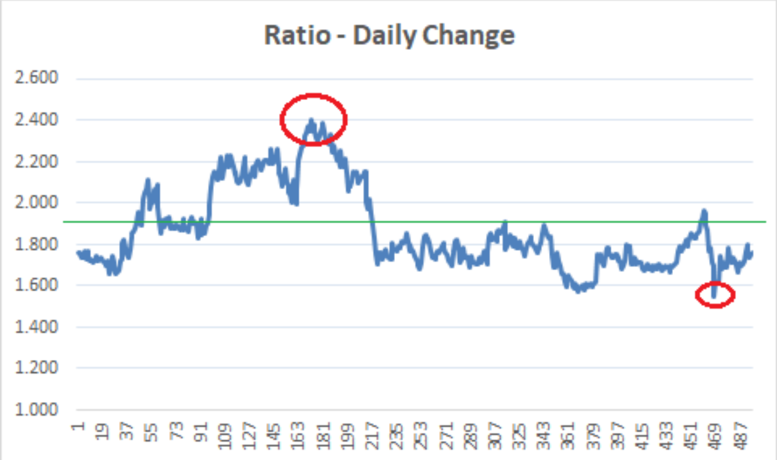
6.3 – The trade trigger

As the name suggests, the pair consists of two stock. Until now, we have not defined how to buy or sell a pair, we will do that later in this chapter. For now, assume that you can buy or sell a pair just like the way you can buy or sell a single stock.

As you may have guessed, the decision to buy or sell a pair is dependent on the variable that you track and the variable itself could be the spread, differential, or ratio. For the purpose of this discussion, we are going ahead with the Ratio.

Think about it this way – the stock prices change every day, therefore the ratio of the pair itself changes every day. On most of the days, the daily change in the ratio falls within the expected range. However, there could be days when the daily change goes beyond the expected range. These are the days when a pair trading opportunity arises.

Have a look at the chart below –



Casual eyeballing reveals two obvious information –

1. The ratio chart hovers around 1.8 and 2 – probably the ratio's mean is around this price. I've highlighted this with a green line. I'd suggest you check the mean value of the ratio we calculated in the earlier chapters.
2. On most of the days, the ratio hovers above or below the mean value

I want you to pause here and think about this. This is the tipping point in Pair trading, if you can understand everything we have discussed up to now, then the rest is a cakewalk.

The ratio itself is a variable which is derived by dividing stock 1 over stock 2. The ratio changes every day since the stock prices change every day. If you plot the chart of the daily change in the ratio you will notice that the ratio has an average (mean) value and the ratio trades above and below the mean value. Irrespective of where the ratio is today (i.e.

either above or below the mean) – there is a great chance that ratio will come back to mean over the next few days. Notice, I use the word 'great chance', here. This means, that we should be able to quantify the probability of the ratio reverting to mean.

In fact, this phenomenon is referred to as 'Mean reversion' or reversion to mean.

I've circled (in red) two points in the chart where the ratio has deviated away from the mean. The first circle from the left indicates a point where the ratio has deviated higher than the mean value. The 2nd circle from the left indicates a point where the ratio has deviated below the mean value. In both these cases, eventually, the ratio reverted to mean.

Now, if you look at it in another way – we now seem to have an opinion on the direction in which the ratio is likely will move. For example, the first circle where the ratio has moved above the average indicates that the ratio is likely to retrace back to mean. Or in other words, you can short the ratio at the high point and buy it back around the mean. Likewise, the second circle points to an opportunity where one can buy the ratio, with an expectation that the ratio will move back to the average value.

Think about the ratio as a stock or futures. Since the directional movement of the ratio is predictable, we can as well place bets on the directional movement of the ratio itself.

I hope you are getting the point here.

The ratio's value with respect to the mean acts as a key trigger to initiate the trade. If the ratio is –

- Above the mean, the expectation is that the ratio will revert to mean, hence short the ratio
- Below the mean, the expectation is that the ratio will scale back to the mean and hence go long on the ratio

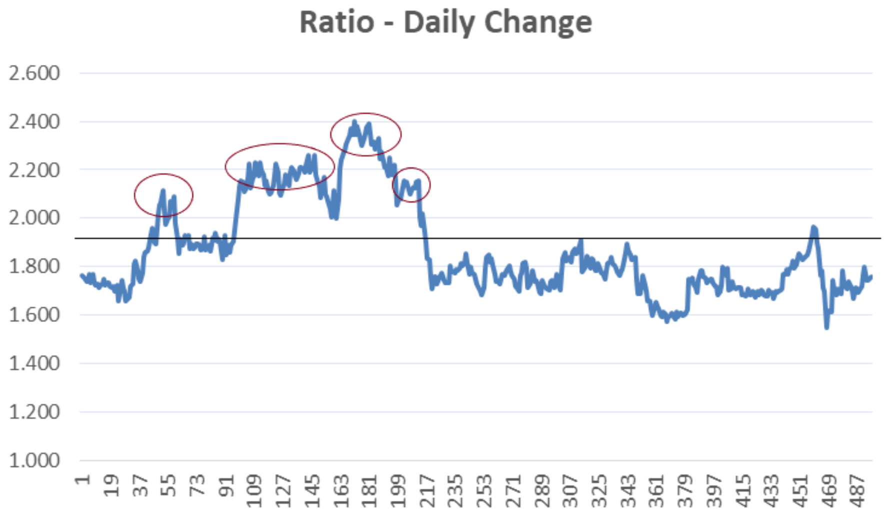
Alright – so far so good. Here are few questions though –

1. The ratio is always above or below the mean value – does this imply there is always a trading opportunity?
2. There are multiple points where the ratio seemed to have bottomed out or peaked, how do we know the exact point at which the trade has to be initiated?

The answers to these questions lie in something called as the 'Density Curve'. Let's figure that out.

6.3 – The Density Curve

Have a look at the chart below –



I've highlighted 4 points on the chart, at all these points, the ratio has traded above the mean. Assume, you were looking at this chart around the time the first circle is marked. Now, just because the ratio has shot up above the mean, would you take the trade? In fact, the same question can be asked every time the ratio has traded above (or below) the mean.

I'm sure you'd agree that this would be a great idea. We need to observe the ratio closely and initiate a trade only when the chance of mean reversion is very high. Or in other words, we need to initiate a trade only when we are reasonably certain that the ratio will slide down to the mean value, **as quickly as possible**.

To put the point across – this is pretty much like a tiger waiting in the ambush to hunt down a prey. Just because the prey is in the open, the tiger will not jump and ruin its chances of a kill. It will attack only when it is convinced that the effort will lead to a kill.

So how do we stay in the ambush and wait for our chance for the kill?

Well, we seek refuge in the good old Normal distribution and its properties. I'm hoping you are aware of normal distribution and its properties by now. Here is a quick recap, I'd suggest you read the complete theory, I've discussed this across various chapters in Varsity –

- Within the 1st standard deviation (SD) one can observe 68% of the data
- Within the 2nd standard deviation one can observe 95% of the data
- Within the 3rd standard deviation one can observe 99.7% of the data

So here is what this means with respect to the ratio –

- The ratio, irrespective of where it stands with reference to the mean, has a standard deviation value. For example – it could be just a few points away from the mean and this could translate to say, 0.5 standard deviations from mean

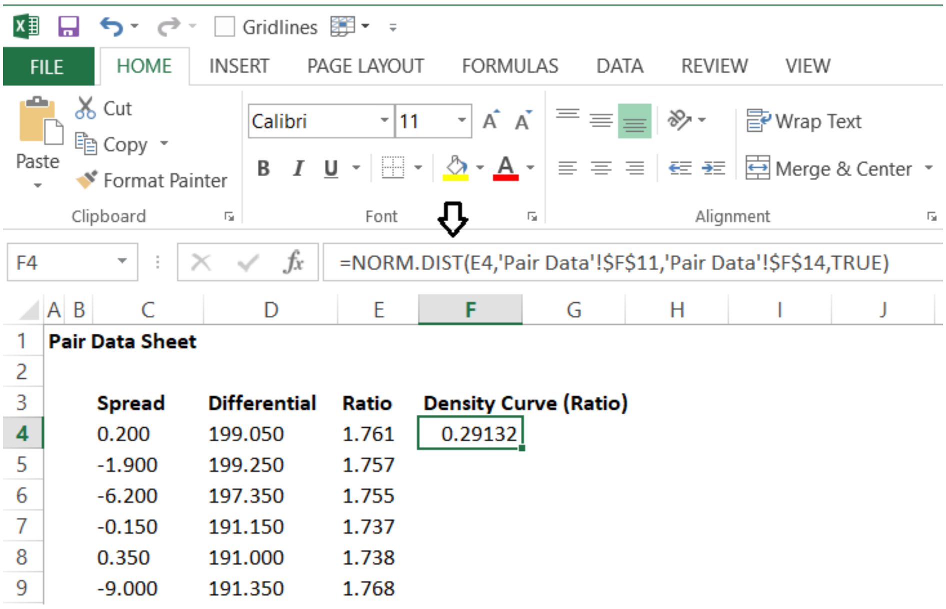
- If the ratio deviates to the 2nd standard deviation, then according to the normal distribution properties, there is only 5% chance of it going higher or in a very loose sense, it poses a 95% chance of reverting to mean.
- Likewise, if the ratio deviates to the 3rd standard deviation, then it only has a 0.3% chance of drifting higher or in a very loose sense, it poses a 99.7% chance of reverting to mean

So at every SD, we can estimate the likelihood of the ratio reverting to mean. This means we can filter out opportunities and initiate a trade only at points where the likelihood of success is high.

This further leads to an interesting take – the key trigger to initiate a trade is not just based on where the ratio is, but also depended on its standard deviation. Given this, it makes sense to directly track the daily standard deviation of the ratio as opposed to the ratio itself.

This can be achieved by tracking the 'Density Curve' of the ratio. The density curve is a non-negative value which lies anywhere between 0 and 1. I'd suggest you **watch this video** on Khan Academy to learn more about **Density Curve.**

Calculating the density curve on excel is quite straightforward. Here is how you can do this, have a look at the image below –



F4

=NORM.DIST(E4,'Pair Data'!$F$11,'Pair Data'!$F$14,TRUE)

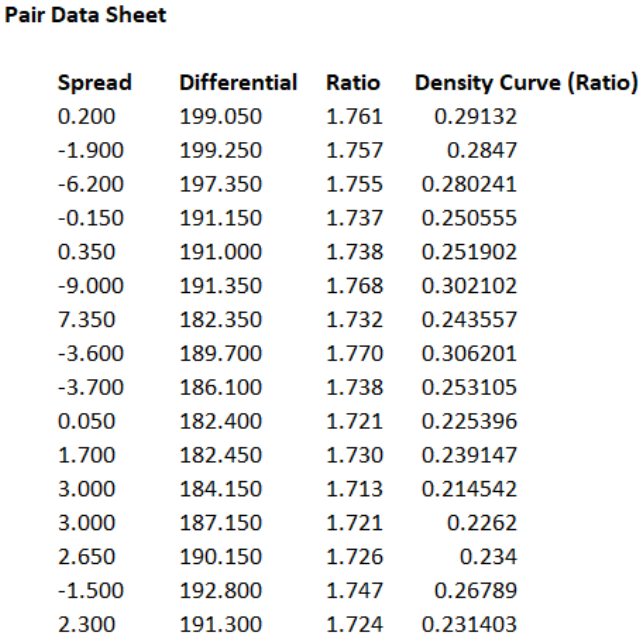
Pair Data Sheet

Spread	Differential	Ratio	Density Curve (Ratio)
0.200	199.050	1.761	0.29132
-1.900	199.250	1.757	
-6.200	197.350	1.755	
-0.150	191.150	1.737	
0.350	191.000	1.738	
-9.000	191.350	1.768	

You can use the inbuilt excel function called Norm.dist for this. The function requires 4 inputs –

- X – this is the daily ratio value
- Mean – this is the mean or average value of the ratio
- Standard Deviation – this is the standard deviation of the ratio
- Cumulative – You have to select true or false, select the default value as true.

I've calculated the density curve value for all variables, here is how the table looks –



Pair Data Sheet

Spread	Differential	Ratio	Density Curve (Ratio)
0.200	199.050	1.761	0.29132
-1.900	199.250	1.757	0.2847
-6.200	197.350	1.755	0.280241
-0.150	191.150	1.737	0.250555
0.350	191.000	1.738	0.251902
-9.000	191.350	1.768	0.302102
7.350	182.350	1.732	0.243557
-3.600	189.700	1.770	0.306201
-3.700	186.100	1.738	0.253105
0.050	182.400	1.721	0.225396
1.700	182.450	1.730	0.239147
3.000	184.150	1.713	0.214542
3.000	187.150	1.721	0.2262
2.650	190.150	1.726	0.234
-1.500	192.800	1.747	0.26789
2.300	191.300	1.724	0.231403

I guess we could break this chapter at this point. In the next chapter, we will look into details on how we can use the density curve to trigger long and short pair trade.

Download the excel sheet used in this chapter.

Key takeaways from this chapter

- Ratio as a variable is more versatile as it captures the valuation elements of the stock

- The ratio tends to trade above or below its mean value
- The idea is the ratio, when it deviates away from the mean, will also tend to revert to mean
- At every point at which the ratio deviates, we can measure the probability of its reversion to mean
- The above point can be measured by normal distribution
- The density curve is a non-negative value which varies between 0 and 1. This can be easily calculated on MS Excel by using an in build function.

CHAPTER 7

PTM1, C5 – The Pair Trade

7.1 – Quick Reminder

We closed the previous chapter with a note on Density curve and how the value of the density curve helps us spot pair trading opportunity. In this chapter, we will work towards identifying and initiating an actual trade and learning other dynamics associated with a pair trade.

Just as a reminder – the techniques we have discussed so far in pair trading (i.e. from chapter 1 through 7) is from the book called 'Trading Pair', by Mark Whistler. The good part about this technique is the simplicity and the part that I'm not too conformable with this technique is also its simplicity. Over time I've improved technique to pair trade, which I will discuss from the next chapter onwards.

Why not discuss the 2nd method directly, you may ask – well, this is because I think Mark Whistler method to pair trade lays an excellent foundation and it helps understand the slightly more complex pair trading technique better. So let me attempt to finish the Mark Whistler's method in this chapter and move to the next method to pair trade.

Now, because I'll discuss this other technique to pair trade, I'll take the liberty to not really get into the nuances of the trade set up. I'll instead focus on the broad trade set up.

So let's get started on it.

7.2 – Digging into Density curve

The density curve acts as a key trigger for us to identify an opportunity to trade. I want you to pay attention to the following two things –

1. The density curve is calculated based on the time series data, and the time series data in our context is the 'ratio' – as you may recall from the previous chapter, the main inputs to calculate the density curve is the ratio's time series, the ratio's mean, and the ratio's standard deviation
2. The density curve is a value – varying between 1 and 0. The value of the density curve helps us understand the probability of the ratio, falling back to the mean.

I understand the 2nd statement may confuse some of the readers, but at this point, I'd suggest you keep this statement in mind. You will understand what I mean by this as we proceed.

Let us spend a little time on the normal distribution, I know we have discussed this multiple times in the past, but bear with me one more time.

The time series data (like the ratio) typically have an average (or mean) value. For example, the average value for the ratio time series is 1.87 (we calculated this in the earlier chapter). More often than not, the value of the ratio tends to lie around the mean value. If the value of the ratio drifts away from the mean, then one can expect the value of the ratio to gravitate back to the mean.

For example, if the latest value of the ratio shoots up to 2.5, then over time, one can expect the value of the ratio to fall to 1.87 and likewise if the value of the ratio plummets.

Now here is a question – If the ratio drifts away from the mean (which is bound to happen on a daily basis), is there a way wherein we can quantify the probability of the ratio to move back to the mean, again?

For example, if the latest ratio value is at 2.5, we all know it will fall to a mean of 1.87, but what is the probability of this occurring? Is it 10%, 20% or 90%?

This is where the density curve comes in handy. The value of the density curve tells us how far, in terms of standard deviation, the ratio has deviated away from its mean. Now, if the value is in terms of standard deviation, then naturally there is a probability assigned to it, and eventually, this probability helps us set up a trade.

Let me give you a quick example.

Consider the following data –

Latest ratio – 2.87

Ratio Mean – 1.87

Density curve – 0.92

Here is how you will interpret this data – the 0.92 value of the density curve indicates that the latest ratio of 2.87 has approximately deviated to the 2nd standard deviation and there is approximately 95% chance that the ratio of 2.87 will fall back to its average value of 1.87.

How did we arrive at this? I mean what tells us that the ratio of 2.87 is approximately near the 2nd standard deviation? Well, we infer this by looking at the corresponding density curve value i.e. 0.92.

The density curve value from 0 to 1 represents the standard deviation values. For example –

1. The density curve of 0.16 implies that the corresponding value is at the -1 standard deviation below the mean
2. The density curve value of 0.84 implies that the corresponding value is at the +1 standard deviation above the mean
3. The density curve value of 0.997 implies that the corresponding value is at the 3 standard deviations above the mean

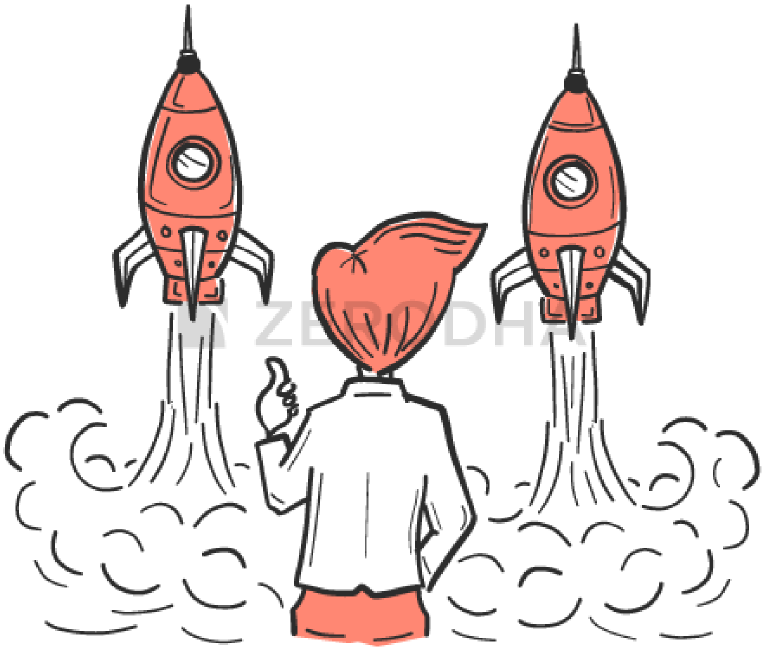
Once I know the standard deviation, I'll also know the probability.

But How did I arrive at 0.16, 0.84, 0.997 etc. in the first place? Well, these are standard deviation values, I will skip dwelling further into standard deviation, instead give you a table which you can use as a ready reckoner –

Density Curve value	How many Standard deviation away	Probability of reverting to mean
0.16	– 1 SD	65%
0.025	– 2 SD	95%
0.003	– 3 SD	99.7%
0.84	+ 1 SD	65%
0.974	+ 2 SD	95%
0.997	+ 3 SD	99.7%

Given the above, if I see the density curve value of around 0.19, I know the ratio is around the – 1ststandard deviation, hence the probability of the ratio to move back to mean is around 65%. Or if the density curve value is around 0.999, I know the value is around the – 3SD, hence the probability of the ratio to move back to mean is around 99.7%

So on and so forth.



7.3 – The first pair trade

So, finally, here we are, very close to showcasing our first Pair trade. Few points to remember –

1. The ratio is calculated by dividing Stock A over Stock B. In our example, Stock A is Axis Bank and Stock B is ICICI Bank. So Ratio = Axis Bank / ICICI Bank
2. The ratio value changes daily, based on the stock prices of Axis Bank and ICICI Bank
3. The ratio and its corresponding density curve value has to be calculated daily

The trading philosophy is as below –

1. If two business are alike and operate in the same landscape – like Axis Bank and ICICI Bank, then their stock prices tend to move together
2. Any change in the business landscape will affect the stock prices of both the companies
3. A stray incident can cause the stock price of one company to deviate away from the stock price of the other. On such days, the ratio to deviates

4. We look for such deviations to identify good trading opportunities

So essentially, a pair trader tracks the ratio and its corresponding density curve value. A pair trade is set up when the ratio (and the density curve) has deviated convincingly enough from the mean value.

This leads us to the next obvious question – what is convincingly enough? Or in other words, at what value of the density curve, should we initiate the trade?

Here is a general guideline to set up a pair trade –

Trade Type	Trigger (density curve)	Standard Deviation	Target	Stoploss
Long	Between 0.025 & 0.003	Between 2nd & 3rd	0.25 or lower	0.003 or higher
Short	Between 0.975& 0.997	Between 2nd & 3rd	0.975 or lower	0.997 or higher

The idea is to initiate a trade (either long or short) when the ratio is between 2nd and 3rd standard deviation and square off the position as it goes below the 2nd standard deviation. Obviously, the closer it goes toward the mean, the higher is your profit.

Let's set up a trade based on the above table, for this, I'd suggest you **download** the excel sheet available towards the end of the previous chapter.

On 25th Oct 2017, the density curve value was 0.05234 and the corresponding ratio value was 1.54. This is a decent **long pair** trade set up. Although this does not fall within the preview of a long trade (we need the density curve to be between 0.025 and 0.003), I guess this is the best value in the time series we are considering.

If the ratio is defined as Stock A / Stock B, then –

1. A long trade requires you to buy Stock A and Sell Stock B
2. A short trade requires you to sell Stock A and Buy Stock B

We have defined the ratio as Axis / ICIC, hence, on 25th closing, one would –

1. Buy Axis Bank @ Rs.473
2. Sell ICICI Bank @ 305.7

The lot size for Axis is 1200, hence the contract value is 1200 * 473 = Rs. 567,600/-. The lot size of ICICI Bank is 2750, hence the contract value is Rs. 840,675/-.

Ideally, we need to stay long and short of the same Rupee value. This is also called 'Rupee Neutrality', but I'll skip this part for now. We will take the concept of Rupee neutrality to a different dimension when we take up the next pair trading technique.

So, once the trade is set up, we now have to wait for the pair to move towards the mean. Ideally, the best pair trade is when you initiate a trade near the 3rd SD and wait for the ratio to move to the mean, but then this could happen over a long period, and the mark to market could be quite painful. In the absence of deep pockets to accommodate for mark to market, one has to be quick in closing a pair trade.

On 31st Oct 2017, the ratio moved up to 1.743 and the corresponding density curve value was 0.26103, which is roughly the target density curve value. Hence once can consider closing the trade.

We Sell Axis Bank @ 523 and buy back ICIC at 300.1. The P&L and other details are as follows –

Date	Stock	Trade	Lot Size	Sq off date	Sq off Price	P&L
25th Oct	Axis Bank	Buy @ 473	1200	31st Oct	Sell @ 523	50*1200 = 60K

25th Oct	ICICI Bank	Sell @ 305.7	2750	31st Oct	Buy @300.1	5.6*2750 =15.4K
Total P&L						**Rs.75,400/-**

If you notice, the bulk of the profits comes from Axis Bank, this indicates that Axis Bank had deviated away from the regular trading pattern.

Not bad eh?

Let's look at a short trade now.

On 9th August 2016, the density curve printed a value of 0.99063156, close enough to initiate a short pair trade. Remember in a short trade, we sell Axis and buy ICICI.

If you find it confusing to remember which one to buy and sell, think of it this way – the numerator is the dominating stock, so if the pair trade demands you to go long, then buy the numerator. Likewise, if the pair trade is to short, the short the numerator. Whatever you do with the numerator, the opposite trade happens with the denominator.

Hence we sell Axis Bank (numerator) and sell ICICI Bank (denominator).

Trade details are as follows –

- Short Axis @ 574.1
- Buy ICICI @ 245.35
- Ratio – 2.34
- Corresponding Density Curve value – 0.99063156

Once initiated, the opportunity close this trade occurred on 8th Sept, (yes, the trade was held open for almost a month). The trade details were –

- Buy Axis @ 571
- Sell ICICI @ 276.33
- Ratio – 2.27
- Corresponding Density Curve value – 0.979182

Agreed, once could have waited a bit longer to for the density curve to fall further, but then like I said before, the pair trader has to strike a balance between the time and mark to markets.

The P&L for the trade is as below –

Date	Stock	Trade	Lot Size	Sq off date	Sq off Price	P&L
9th Aug	Axis Bank	Sell @ 574.1	1200	8th Sept	Buy @ 571	3.1*1200 = 3.72K
9th Aug	ICICI Bank	Buy @ 245.3	2750	8th Sept	Sell @276.33	31.03*2750 = 85.3K
		Total P&L				**Rs.89,052/-**

Again, the bulk of the profit comes from one of the stocks i.e. ICICI, indicating that ICICI had probably deviated away from its course.

I must confess, both the trades did not really fall under the prescribed table giving you the guideline to enter and exit the pair trade. But like I said before, use the table as a reference and build your expertise around it.

I'd encourage you to look for any other opportunities in the Axis & ICICI Bank example.

I hope the P&L of pair trade is incentivizing you enough to learn more about pair trading. I'll deliberately stop here, to ensure you soak in everything that we have discussed. I'll leave you with few final points.

1. Everything we have learned so far accounts to about 25% of what I intend to discuss going ahead
2. These first 7 chapter discusses a very basic pair trading technique, mainly to help lay a foundation
3. We have not adhered to strict trade definitions – stop loss, targets etc. If you notice, I've kept things quite generic
4. Neutrality of both the positions is a key angle, we have not discussed that yet
5. We are yet to discuss the risk associated with Pair trading
6. Pair trading is a margin money guzzler, so one needs to have sufficient funds to pair trade, but the P&L is worth it
7. For a given pair, at the most 2-3 signals is what you can expect in a year. So one has to track multiple pairs to find continuous opportunities in the market

Anyway, I hope I've managed to ignite your curiosity to learn more on Pair Trading. I'm eager to move forward, I hope you are too!

Download the excel sheet.

Key takeaways from this chapter

1. The density curve acts as a key trigger to initiate a pair trade
2. A pair trade is initiated when the ratio drifts to a value between 2 and 3 standard deviation
3. A pair trade is closed when the ratio approaches the mean
4. Long pair trade requires you to buy the numerator and sell the denominator

5. Short pair trade requires you to sell the numerator and buy the denominator
6. Typically, the bulk of P&L comes from one of the stocks which have deviated away from the regular pair trade
7. Pair trade can be live for an extended period, but the P&L makes the wait worth it
8. Pair trade is a margin money guzzler.

CHAPTER 8

Pair trade Method 2, Chapter 1 (PTM2, C1) – Straight line Equation

8.1 – A straight relationship

Today happens to be 14th of Feb, people around me are excited about Valentine's Day, they are busy celebrating love and relationships. I think Valentine's Day is a packaged affair, meant to boost the revenues of restaurants, jewellers, and gift shops, but then it's just me and my random thoughts.

Anyway, given its valentine's day, I thought it would be a perfect idea to discuss relationships. Don't worry, I'm not going to bore with a clichéd love story or give you any unsolicited advice on maintaining a great relationship, rather I'll talk to you about two sets of numbers and how you can measure the relationship between them if at all there exists one.

In the process, I'll attempt to take you back to your school days, well, at least back to your

high school math class

A quick recap here – Chapter 1 to 7 of this module, we discussed a rather simple technique of pair trading. This was as taught by Mark Whistler. Moving forward from this chapter, we will discuss a slightly more advanced technique of pair trade. This is also called '**Statistical Arbitrage**' or '**Relative value trading**' or RVT in short.

So here we go.

Do you remember the time your math teacher discussed the equation of a **straight line** in the class? If you were like me, you'd have promptly ignored the lecture and looked outside of the window, quietly rebelling against the mainstream education.

But then, if only the teacher had said 'learn this, you'll make money off it someday', the interest level would have been totally different!

Anyway, life always gives you a second chance, so this time around, pay attention, and

hopefully, you will make some money off it

The equation of a straight line reads something like this –

Y = mx + ε

Click here for a detailed explanation, or continue reading for a bare bone explanation.

Before we discuss the equation, a quick note on the notations used –

y = Dependent variable

M = Slope

X = Independent variable

E = Intercept

The equations states, the value of a dependent variable 'y' can be derived from an independent variable 'x', by multiplying x by its slope with y' and adding the intercept 'e' to this product.

Sounds confusing? I guess so

Let me elaborate on this and by the way before you start thinking why we are discussing the straight line equation instead of relative value trading (RVT), then please be rest assured, this concept has deep relevance to RVT!

Consider two fitness freaks, let's call them FF1 and FF2, between the two, FF2 is the kind of guy who wants to go that step extra and something more than what FF1 does. So if FF1 does 5 pushups, FF2 does 10. If FF1 does 20 pull-ups, then FF2 does 40. So on and so forth.   Here is a table on how many pushups they did Monday to Saturday –

Day	FF1	FF2
Monday	30	60
Tuesday	15	30
Wednesday	40	80
Thursday	20	40
Friday	10	20
Saturday	15	???

Now, if you were to guess the number of push-ups FF2 would do on Saturday, what would it be? I guess it's a no-brainer, it would be 30.

This also means – the number of pushups FF2 does, is kind of dependent on the number of pushups FF1 does. FF1 does not really bother about FF2, he will go ahead and do as many pushups his body permits, but FF2, on the other hand, does twice the number of pushup as FF1.

So this makes FF2 a dependent variable and FF1 an independent variable. Or in the straight line equation, FF2 = y and FF1 = x.

FF2 = FF1*M + ε

In simple English, the equation reads like this –

The number of pushups FF2 does is equal to the number of pushups FF1 does, multiplied by a certain number, plus a constant.

That certain number is called the slope (M), which happens to be 2, and the constant or ε happens to be 0. So the equation is –

FF2 = FF1*2 + 0

I hope this is fairly clear now. Let me copy paste the definition I had posted earlier –

The straight line equations states, the value of a dependent variable 'y' can be derived from an independent variable 'x', by multiplying x by its slope with y' and adding the intercept 'e' to this product.

Now, think about another case –

There are two hungry men, let's call them H1 and H2. Just like FF1 and FF2, H2 eats twice the number of paratha as H1 plus 1.5 more. For example, if H1 eats 2 parathas, then H2 will eat 4 plus eat another 1.5. H2 will always ensure he eats that extra 1.5 parathas, no matter how full he is.

So here is the table which gives you count of how many parathas these two hungry men ate over the last 6 days –

Day	H1	H2
Monday	2	5.5
Tuesday	1.5	4.5
Wednesday	1	3.5
Thursday	3	7.5
Friday	3.5	8.5
Saturday	4	???

If you notice, H2 (who is really hungry, all the time), eats twice as much as H1 plus 1.5 paratha extra. So on Saturday, he will eat –

4*2 + 1.5 = 9.5 paratha!

Remember, the number of parathas H2 eats is dependent on how many parathas H1 eats. H1, on the other hand, eats till he is satisfied. Given this, let us a construct a straight line equation for these two hungry men, just like the way we did for the two fitness freaks.

H2 = H1*2 + 1.5

Here, H2 is the dependent variable, whose value is dependent on H1. 2 is the slope, and 1.5 is the constant.

Before we proceed, let’s make a small change in the paratha example, think of ‘Y’ as a diet conscious person. Every day, irrespective of how hungry or full Y is, he eats just 1.5 parathas. Not a morsel more or not morsel less.

So, X eats 3 parathas, Y eats 1.5, X eats 5, Y eats 1.5, X eats 2.5, Y eats 1.5. So on and so forth. So what do you think the equation states?

y = x*0 + 1.5

The slope here is 0, hence, y is not really dependent on x, in fact, the value of y is a constant of 1.5, which is quite obvious. Hopefully, you get the point by now on how you can relate two sets of numbers.

Now forget the fitness, forget the parathas, I’ll give you two sets of random numbers –

X	Y
10	3
12	6
8	4
9	17
20	36
18	22

X is the independent variable and Y is the dependent variable. Given this, do you see a relationship between these two sets of numbers here? Eyeballing the numbers suggest that there is no relationship between X and Y, definitely not like the one which existed in the above two examples. But this does not mean that there is no relationship between the two at all. It's just the relationship is not obvious to the naked eye.

So how do we establish the relationship between the two? To be more precise, how do we figure out the values of the slope' and the constant 'ε'?

Well, say hello to linear regression!

I'll introduce the same to you in the next chapter.

Key takeaways from this chapter

1. A straight line equation can define the relationship between two variables
2. Of the two variables, one of it is dependent and the other one is independent
3. The slope of a straight-line equation, represented by 'm' helps you identify the extent by which the independent variable has to be scaled
4. The term ε represents a constant term
5. If the slope is zero, the Y = ε
6. Sometimes, the relationship between two variables is not obvious
7. When the relationship is not obvious, one can identify the relationship by employing a statistical technique called 'Linear regression'.

CHAPTER 9

PTM2, C2 – Linear Regression

9.1 – Introduction to Linear Regression

The previous chapter laid down a basic understanding of a straight line equation. To keep things simple, we took a very basic example to explain how two variables can be related to each other. Needless to say, the examples were selected in a way that casual eyeballing could reveal the relationship. Towards the end of the chapter we posted a table containing two arrays of numbers – the task was to figure out if there was a relationship between the two sets of numbers, if yes, what how could one express the relationship in the form of a straight line equation. More precisely, what was the intercept and constant?

We will figure how to establish a relationship in this chapter and move closer towards the relative value trading technique. For convenience, let me post the table with the two number arrays once again –

X	Y
10	3
12	6
8	4
9	17

20	36
18	22

Clearly, casual eyeballing does not reveal any information about the relationship between the two sets of numbers. Maybe it does, if you are a mutant, but for a mere mortal like me, it does not work.

Under such circumstances, we rely upon a technique called the ‘Linear Regression’. Linear regression is a statistical operation wherein the input is an array of two sets of numbers and the output contains many different parameters, including the intercept and constant needed for constructing the straight line equation.

To perform the linear regression operation, we will depend on the good old Excel. Here is the step by step guide to perform a simple linear regression on two arrays of numbers. Be

prepared to see a lot of screenshots and instructions

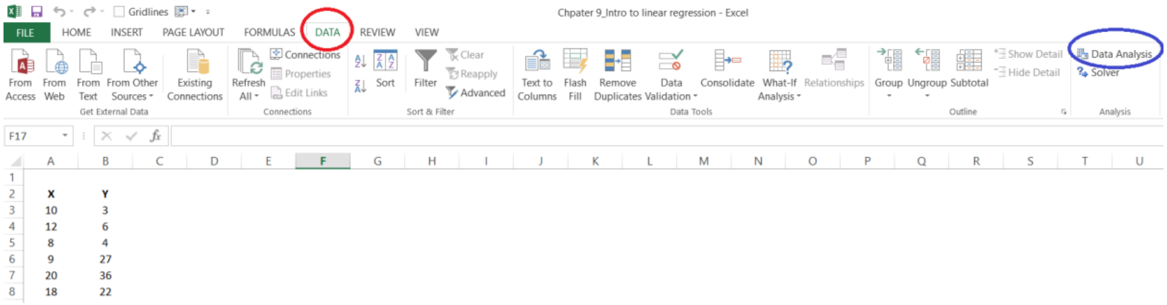
Step 1 – **Install the Plugin**

Open a fresh excel sheet and insert the values of X & Y as seen in the above table. I've done the same as shown below –

X	Y
10	3
12	6
8	4
9	27
20	36
18	22

This is our data set. Do remember, Y is the 'Dependent' variable whose value depends on the independent variable X. Both X and Y will be the input variables for the linear regression operation.

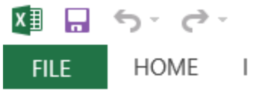
On the excel sheet, click on the Data ribbon as highlighted in red, shown below –



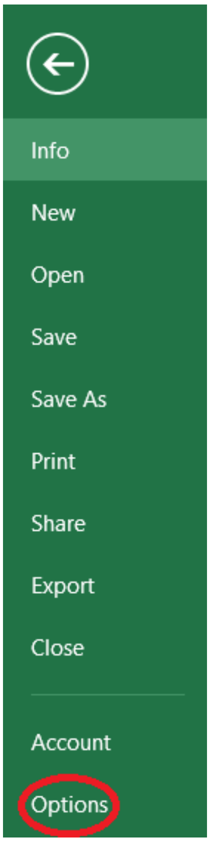
	A	B
1		
2	X	Y
3	10	3
4	12	6
5	8	4
6	9	27
7	20	36
8	18	22

The data ribbon will now show you the 'Data Analysis', option. This is highlighted in blue. Now, some of you may not see this option, if yes, don't panic. I'll tell you what needs to be done.

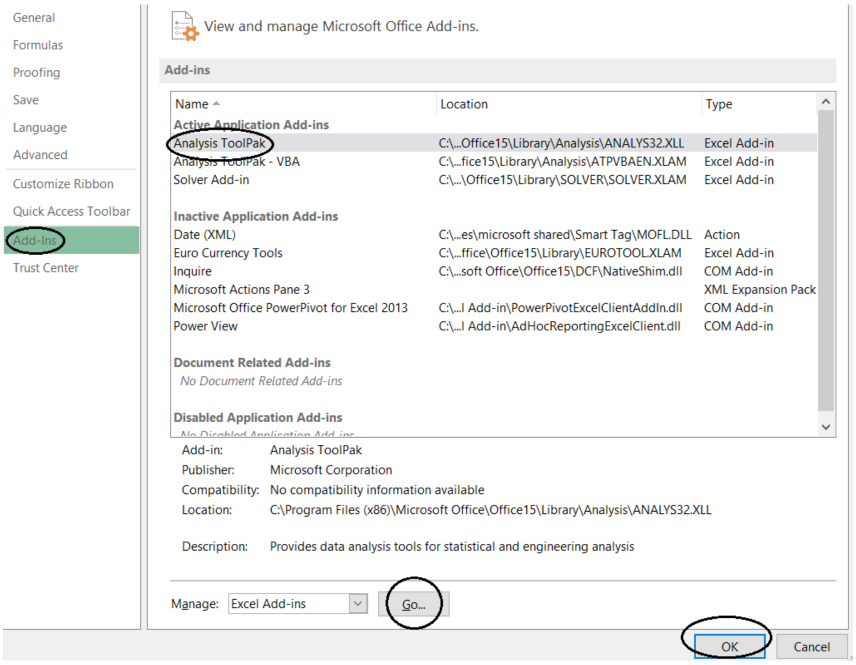
Click on 'File' –



This will open up a new window, and on your left-hand side panel, you will see an option to select 'option' –



Click on the Options, and you will see a bunch of general options to work with. On the left-hand panel, select 'Add-Ins', click on it and then click on the 'Analysis Tool pack'. Then click on 'Go', and finally on 'Ok'. With this, you'd essentially added the 'Data Analysis' option to the data ribbon.



View and manage Microsoft Office Add-ins.

Add-ins

Name	Location	Type
Active Application Add-ins		
Analysis ToolPak	C:\...Office15\Library\Analysis\ANALYS32.XLL	Excel Add-in
Analysis ToolPak - VBA	C:\...fice15\Library\Analysis\ATPVBAEN.XLAM	Excel Add-in
Solver Add-in	C:\...\Office15\Library\SOLVER\SOLVER.XLAM	Excel Add-in
Inactive Application Add-ins		
Date (XML)	C:\...es\microsoft shared\Smart Tag\MOFL.DLL	Action
Euro Currency Tools	C:\...ffice\Office15\Library\EUROTOOL.XLAM	Excel Add-in
Inquire	C:\...soft Office\Office15\DCF\NativeShim.dll	COM Add-in
Microsoft Actions Pane 3		XML Expansion Pack
Microsoft Office PowerPivot for Excel 2013	C:\...l Add-in\PowerPivotExcelClientAddIn.dll	COM Add-in
Power View	C:\...l Add-in\AdHocReportingExcelClient.dll	COM Add-in
Document Related Add-ins		
No Document Related Add-ins		
Disabled Application Add-ins		
No Disabled Application Add-ins		

Add-in: Analysis ToolPak
Publisher: Microsoft Corporation
Compatibility: No compatibility information available
Location: C:\Program Files (x86)\Microsoft Office\Office15\Library\Analysis\ANALYS32.XLL

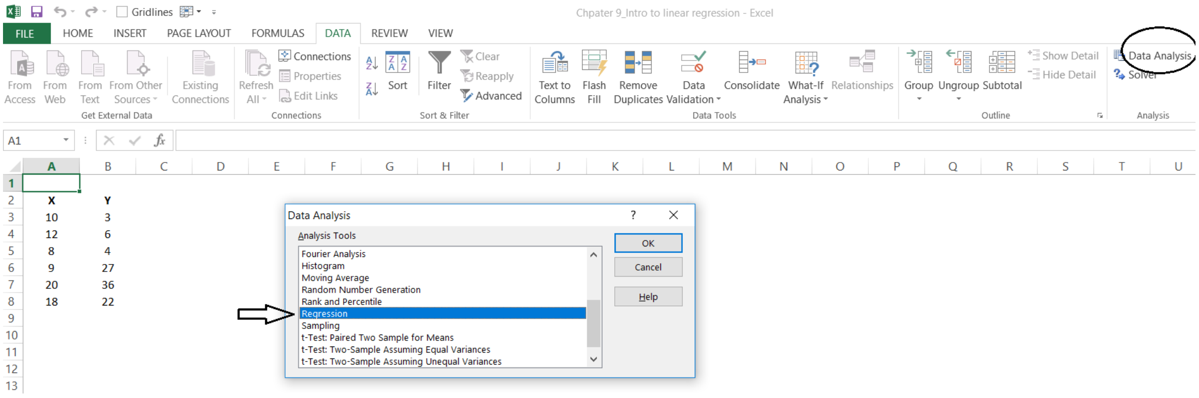
Description: Provides data analysis tools for statistical and engineering analysis

Manage: Excel Add-ins

Close the excel sheet and restart your system and you are good to roll.

Step 2 – Enter the values

So we proceed further based on the assumption that your excel sheet has the data analysis pack. The next step is to invoke the linear regression function within the data analysis pack. To do this, click on the 'Data' ribbon, and select the Data Analysis. This will open up a pop-up, which will have a list of statistical operations which you can perform on data sets. Select the one which says 'Regression'.



Chpater 9_Intro to linear regression - Excel

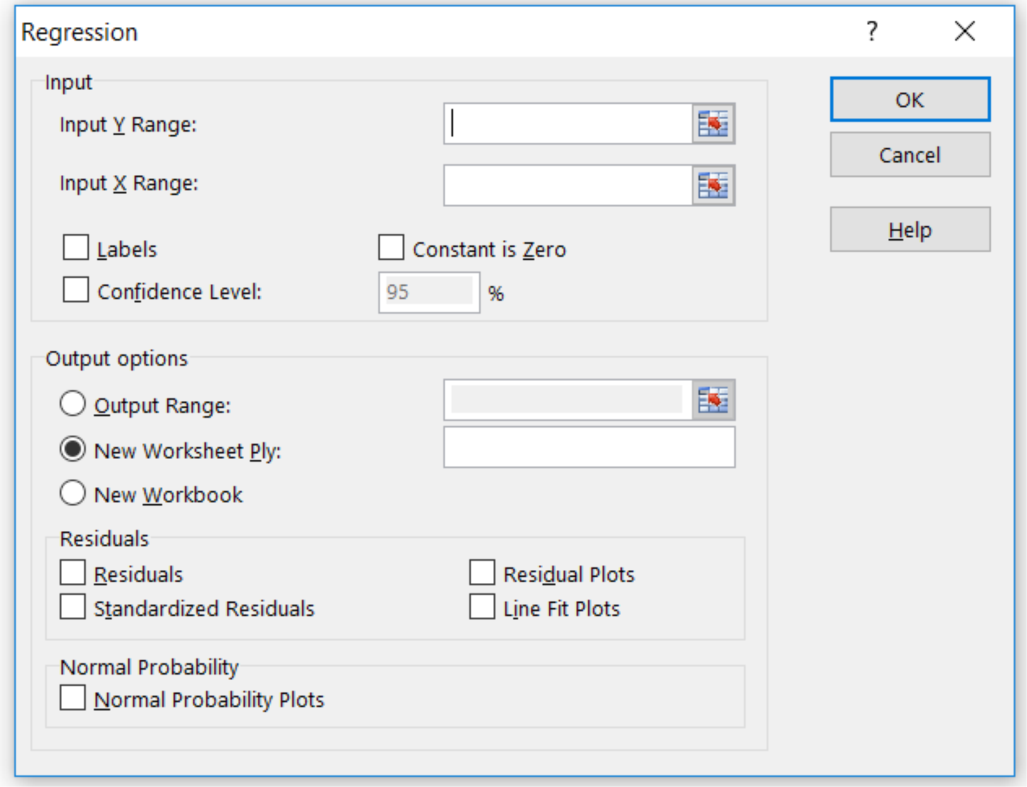
	A	B
1		
2	X	Y
3	10	3
4	12	6
5	8	4
6	9	27
7	20	36
8	18	22

Data Analysis

Analysis Tools

- Fourier Analysis
- Histogram
- Moving Average
- Random Number Generation
- Rank and Percentile
- Regression
- Sampling
- t-Test: Paired Two Sample for Means
- t-Test: Two-Sample Assuming Equal Variances
- t-Test: Two-Sample Assuming Unequal Variances

Select regression and click ok, you will see the following pop up –



Regression

Input

Input Y Range:

Input X Range:

□ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

□ Residuals □ Residual Plots

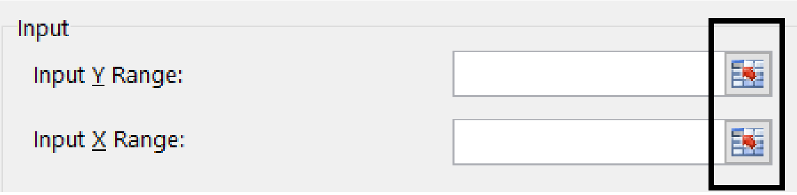
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

As you can see, there are a bunch of fields here. I'd suggest you pay attention to the first section, which is the input section. There are two fields here – 'Input Y Range' and 'Input X Range'. As you may have imagined, Y is for the dependent variable and X is for the dependent variable.

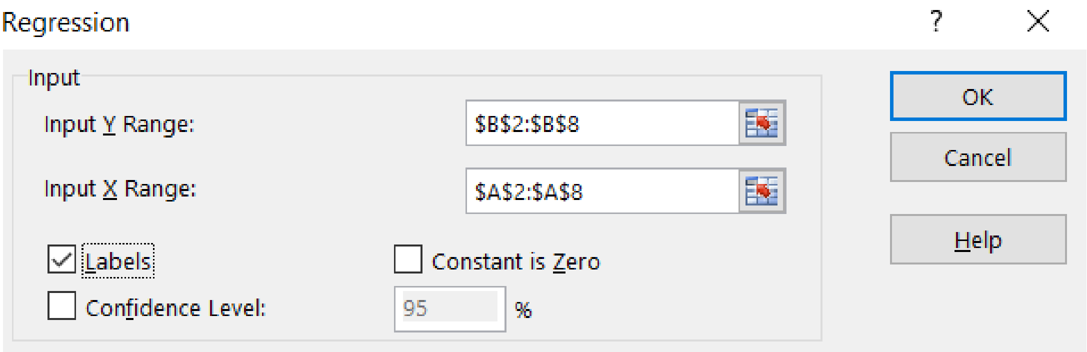
This is where we feed in the X and Y series data. To do that, click on the input channel and select Y and X range –



Input

Input Y Range:

Input X Range:



Regression

Input

Input Y Range: $B$2:$B$8

Input X Range: $A$2:$A$8

☑ Labels

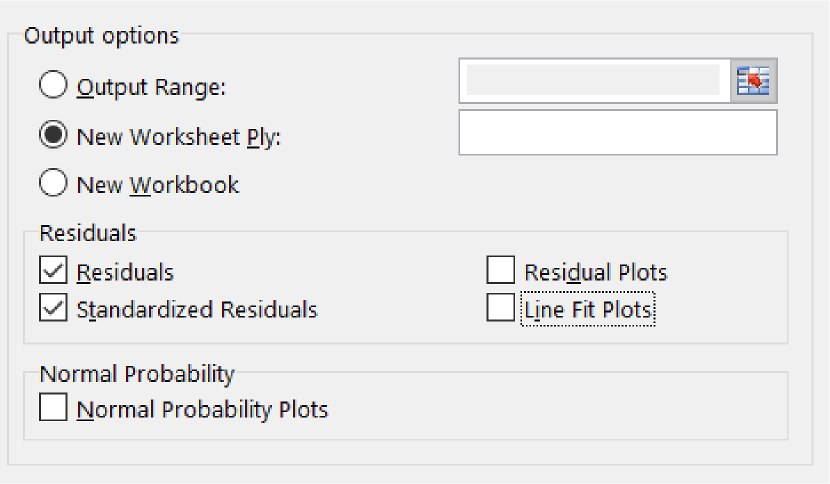
□ Constant is Zero

□ Confidence Level: 95 %

Also, please notice that I've checked the label box, this indicates that the first cell value i.e. A2 and B2 contain the series label i.e. X & Y respectively.

I'd suggest you ignore the other input values for now.

On the output side, ensure you've clicked the following –



Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals

☑ Standardized Residuals

□ Residual Plots

□ Line Fit Plots

Normal Probability

□ Normal Probability Plots

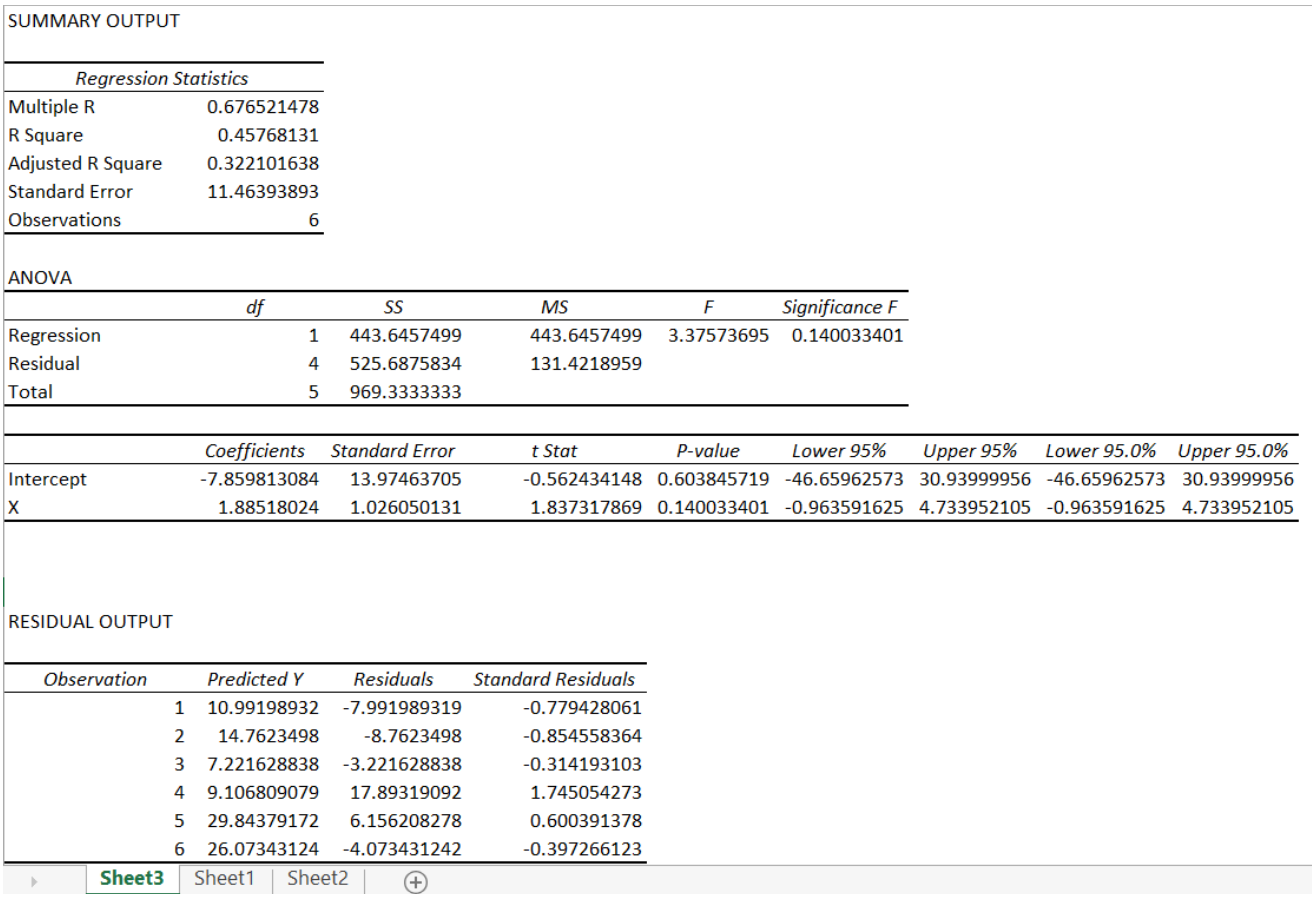
Selecting ‘New worksheet’, ensures that the output data is printed on a new worksheet. I’ve also clicked on two other variables called – Residuals and Standardized Residuals. I will talk about these two variables at a later point. For now, just ensure they are selected.

With this, you are good to perform the linear regression operation. Click on the ‘Ok’ button which is available in the right-hand top corner.

Excel will now take these inputs and perform the linear regression operation, the results will be posted in a new sheet within the same workbook.

9.2 – Linear Regression Output

So here is how the linear regression output looks and as expected, the summary of the output is presented in a new sheet.



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

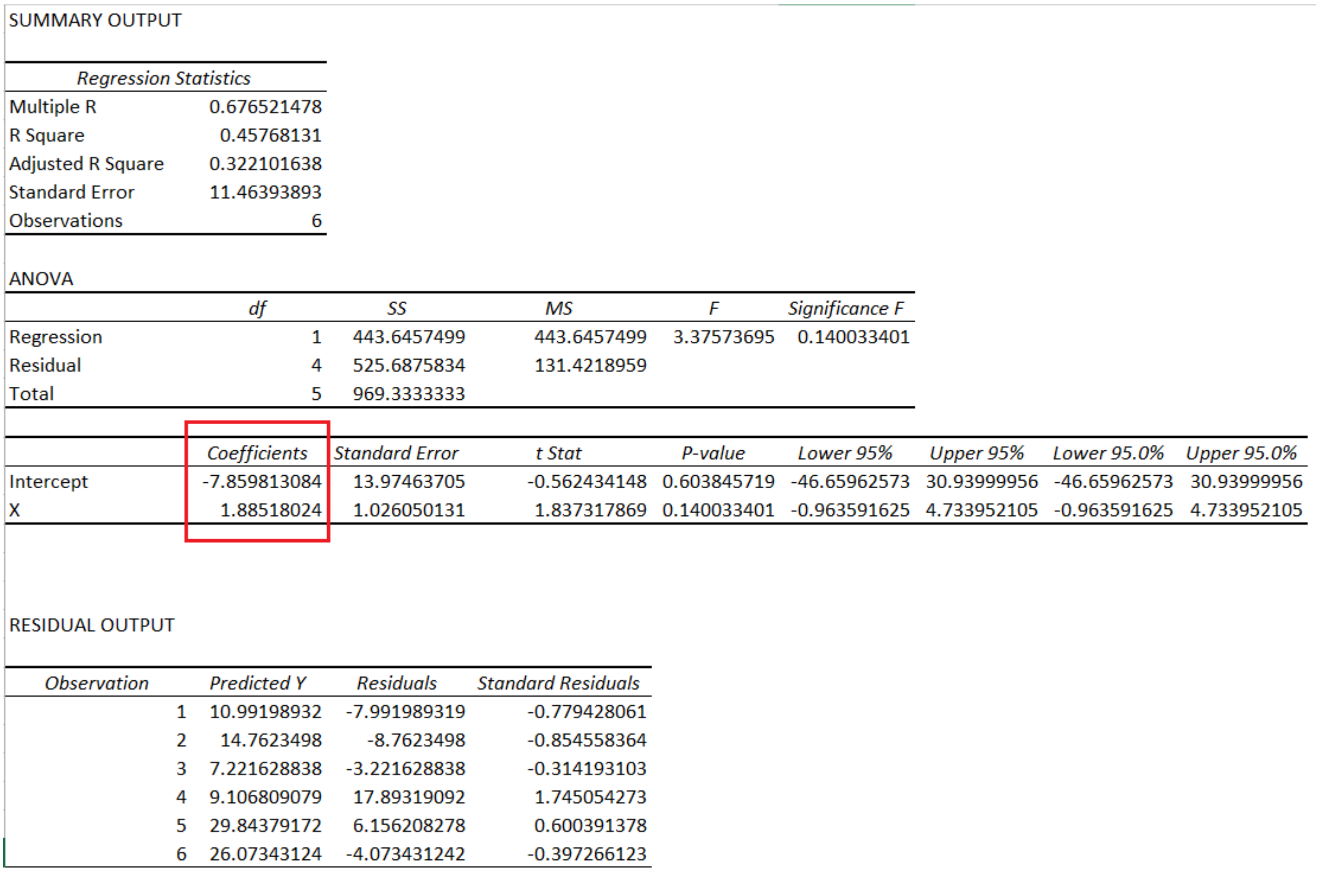
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Standard Residuals
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

Agreed, the summary output is quite scary at the first glance. It has lots and lots of information. We will unravel this output in bits and pieces as we proceed.

For now, let’s concentrate on finding our slope and intercept. I’ve highlighted this for you in the below snapshot –



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	*Predicted Y*	*Residuals*	*Standard Residuals*
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

The data points highlighted in red contains the coefficients we are looking for i.e. the intercept (or constant) and the slope (denoted by x).

Some of you may be confused with the slope being represented by x, I understand its misleading, it would have been best if it was M instead of x as it would match the straight-line equation, but then I guess we will have to live with x for slope.

So,

- Slope of the equation = 1.885
- Intercept (or constant) = -7.859813.

Given this, the straight-line equation for the arbitrary set of data is –

y = 1.885*x + (-7.859813) or

y = 1.885*x – 7.859813

So what does this really mean?

Well, if you recollect from the previous chapter, this equation essentially helps us predict the value of y or the dependent variable for a certain x. Let me repost the table here for the sake of convenience –

X	Y
10	3
12	6
8	4
9	17
20	36
18	22
15	**??**

I've added a new data point for x here i.e. 15, now using the slope and intercept, we can predict the value of y. Let's do that –

y = 1.885 * 15 – 7.859813

= 28.275 – 7.859813

= **20.415**

So, if x is 15, then most likely, the predicted value of y is 20.415.

How accurate is this prediction, you may ask?

Well, it's not accurate. It is only an estimation. For example, consider the value of x is 18 (refer to the last but one data point), then according to the straight line equation, the value of y should be –

y = 1.885*18 – 7.859813

= 33.93 – 7.859813

= 26.07019

However, the actual value of y is 22.

This leads us two values of y –

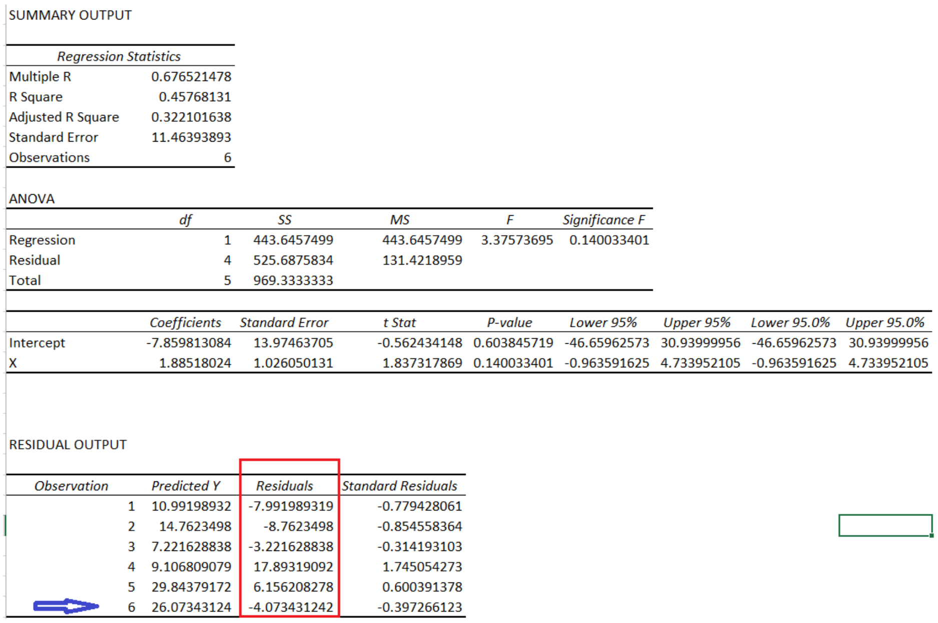
1. Predicted value of y via the straight line equation
2. Actual value of y

The difference between the two values of y is called **the residuals**. For example, the residual for y (difference between actual and predicted y), when x = 18 is

26.07019 – 22

= **4.070187**

The summary output when you perform linear regression also contains the residuals, I've highlighted the same in the snapshot below –



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.676521478
R Square	0.45768131
Adjusted R Square	0.322101638
Standard Error	11.46393893
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	443.6457499	443.6457499	3.37573695	0.140033401
Residual	4	525.6875834	131.4218959		
Total	5	969.3333333			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-7.859813084	13.97463705	-0.562434148	0.603845719	-46.65962573	30.93999956	-46.65962573	30.93999956
X	1.88518024	1.026050131	1.837317869	0.140033401	-0.963591625	4.733952105	-0.963591625	4.733952105

RESIDUAL OUTPUT

Observation	Predicted Y	Residuals	Standard Residuals
1	10.99198932	-7.991989319	-0.779428061
2	14.7623498	-8.7623498	-0.854558364
3	7.221628838	-3.221628838	-0.314193103
4	9.106809079	17.89319092	1.745054273
5	29.84379172	6.156208278	0.600391378
6	26.07343124	-4.073431242	-0.397266123

I've also highlighted the residual when x = 18, which is what we calculated above.

To give you a heads up – the bulk of the focus for carrying out the relative value trade depends on the residuals. Stay tuned!

Download the excel sheet **here**.

Key takeaways from this chapter

1. Linear regression is a statistical operation which helps you construct a straight line equation
2. Linear regression can be performed on excel. One needs to install the excel plugin to perform linear regression
3. Amongst many other output variables, linear regression gives out the values of the slope and intercept
4. With the help of the slope and intercept, one can predict the value of y
5. The difference between actual y and predicted y is called the residual
6. The residual is also a part of the output summary

CHAPTER 10

PTM2, C3 – The Error Ratio

10.1 – Who is X and who is Y?

I hope the previous chapter gave you a basic understanding of linear regression and how one can conduct the linear regression operation on two sets of data, on MS Excel. Remember, we are talking about two variables here – X and Y.

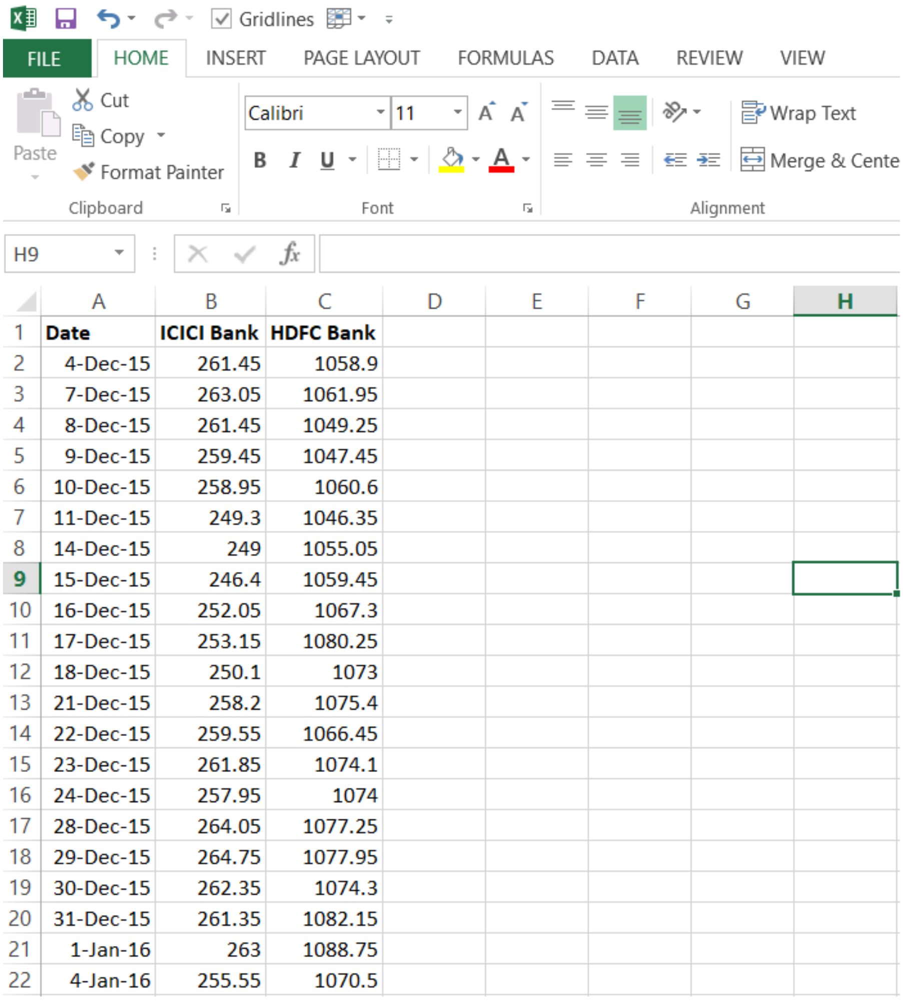
X is defined as the independent variable and Y is the dependent variable. If you've spent time thinking about this, then I'm certain you'd have guessed X and Y will eventually be two different stocks.

In fact, let us just go ahead and run a linear regression on two stocks – maybe HDFC Bank and ICICI Bank and see what results we get.

I'm setting ICICI Bank as X and HDFC Bank as Y. A quick note on data before we proceed –

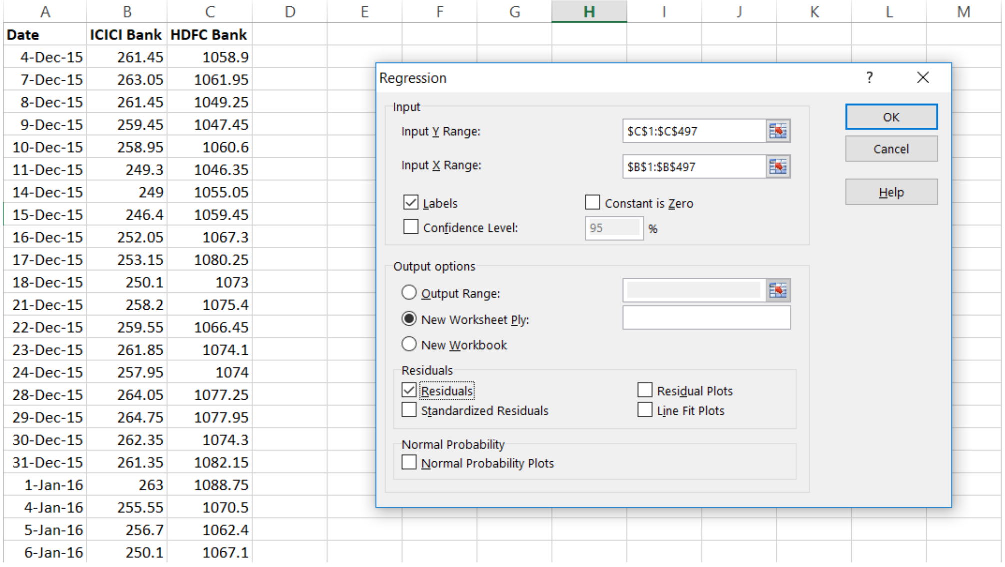
1. Make sure your data is clean – adjusted for splits, bonuses, and any other corporate actions
2. Make sure the data matches the exact dates – for instance, the data I have for both the stocks here runs from 4th of Dec 2015 to 4th Dec 2017.

Here is how the data looks –



	A	B	C	D	E	F	G	H
1	Date	ICICI Bank	HDFC Bank					
2	4-Dec-15	261.45	1058.9					
3	7-Dec-15	263.05	1061.95					
4	8-Dec-15	261.45	1049.25					
5	9-Dec-15	259.45	1047.45					
6	10-Dec-15	258.95	1060.6					
7	11-Dec-15	249.3	1046.35					
8	14-Dec-15	249	1055.05					
9	15-Dec-15	246.4	1059.45					
10	16-Dec-15	252.05	1067.3					
11	17-Dec-15	253.15	1080.25					
12	18-Dec-15	250.1	1073					
13	21-Dec-15	258.2	1075.4					
14	22-Dec-15	259.55	1066.45					
15	23-Dec-15	261.85	1074.1					
16	24-Dec-15	257.95	1074					
17	28-Dec-15	264.05	1077.25					
18	29-Dec-15	264.75	1077.95					
19	30-Dec-15	262.35	1074.3					
20	31-Dec-15	261.35	1082.15					
21	1-Jan-16	263	1088.75					
22	4-Jan-16	255.55	1070.5					

I'll run the linear regression on these two stocks (I've explained how to do this in the previous chapter), also do note, I'm running this on the stock prices and not really on stock returns –



A	B	C
Date	ICICI Bank	HDFC Bank
4-Dec-15	261.45	1058.9
7-Dec-15	263.05	1061.95
8-Dec-15	261.45	1049.25
9-Dec-15	259.45	1047.45
10-Dec-15	258.95	1060.6
11-Dec-15	249.3	1046.35
14-Dec-15	249	1055.05
15-Dec-15	246.4	1059.45
16-Dec-15	252.05	1067.3
17-Dec-15	253.15	1080.25
18-Dec-15	250.1	1073
21-Dec-15	258.2	1075.4
22-Dec-15	259.55	1066.45
23-Dec-15	261.85	1074.1
24-Dec-15	257.95	1074
28-Dec-15	264.05	1077.25
29-Dec-15	264.75	1077.95
30-Dec-15	262.35	1074.3
31-Dec-15	261.35	1082.15
1-Jan-16	263	1088.75
4-Jan-16	255.55	1070.5
5-Jan-16	256.7	1062.4
6-Jan-16	250.1	1067.1

Regression

Input

Input Y Range: $C$1:$C$497

Input X Range: $B$1:$B$497

☑ Labels ☐ Constant is Zero

☐ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

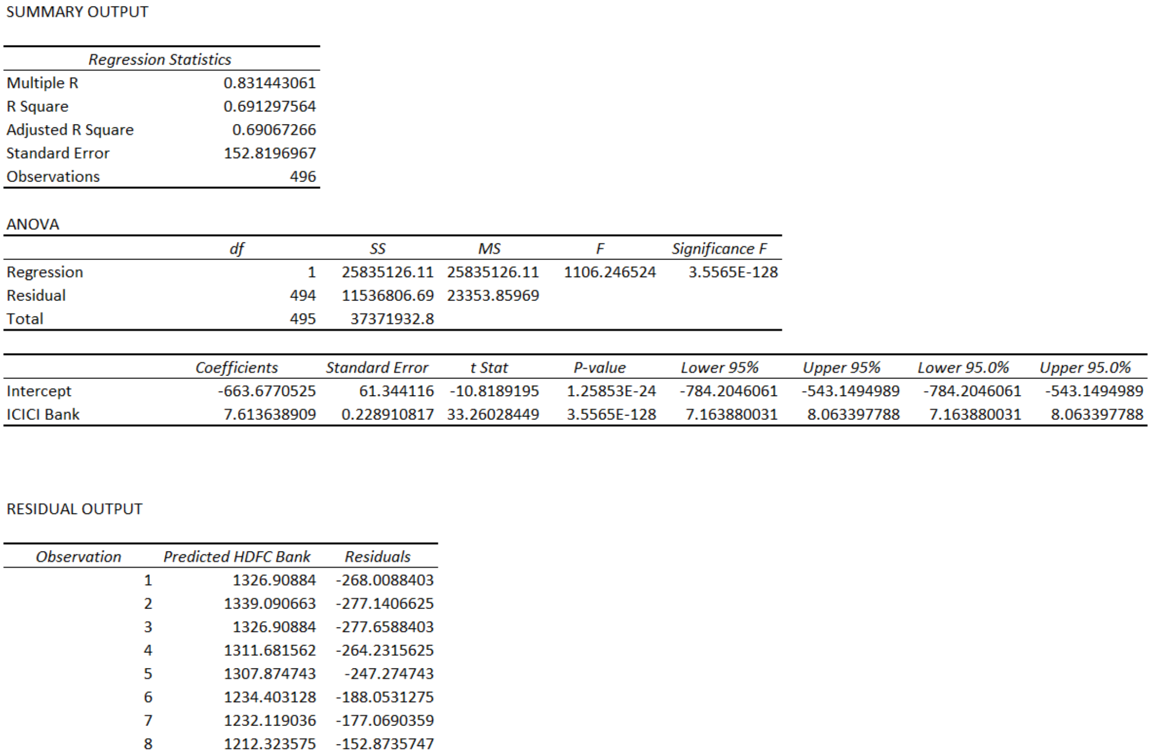
☑ Residuals ☐ Residual Plots

☐ Standardized Residuals ☐ Line Fit Plots

Normal Probability

☐ Normal Probability Plots

The result of the linear regression is as follows –



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	SS	MS	F	Significance F
Regression	1	25835126.11	25835126.11	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.85969		
Total	495	37371932.8			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-663.6770525	61.344116	-10.8189195	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.26028449	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

RESIDUAL OUTPUT

Observation	Predicted HDFC Bank	Residuals
1	1326.90884	-268.0088403
2	1339.090663	-277.1406625
3	1326.90884	-277.6588403
4	1311.681562	-264.2315625
5	1307.874743	-247.274743
6	1234.403128	-188.0531275
7	1232.119036	-177.0690359
8	1212.323575	-152.8735747

Since ICICI is independent and HDFC is dependent, the equation is –

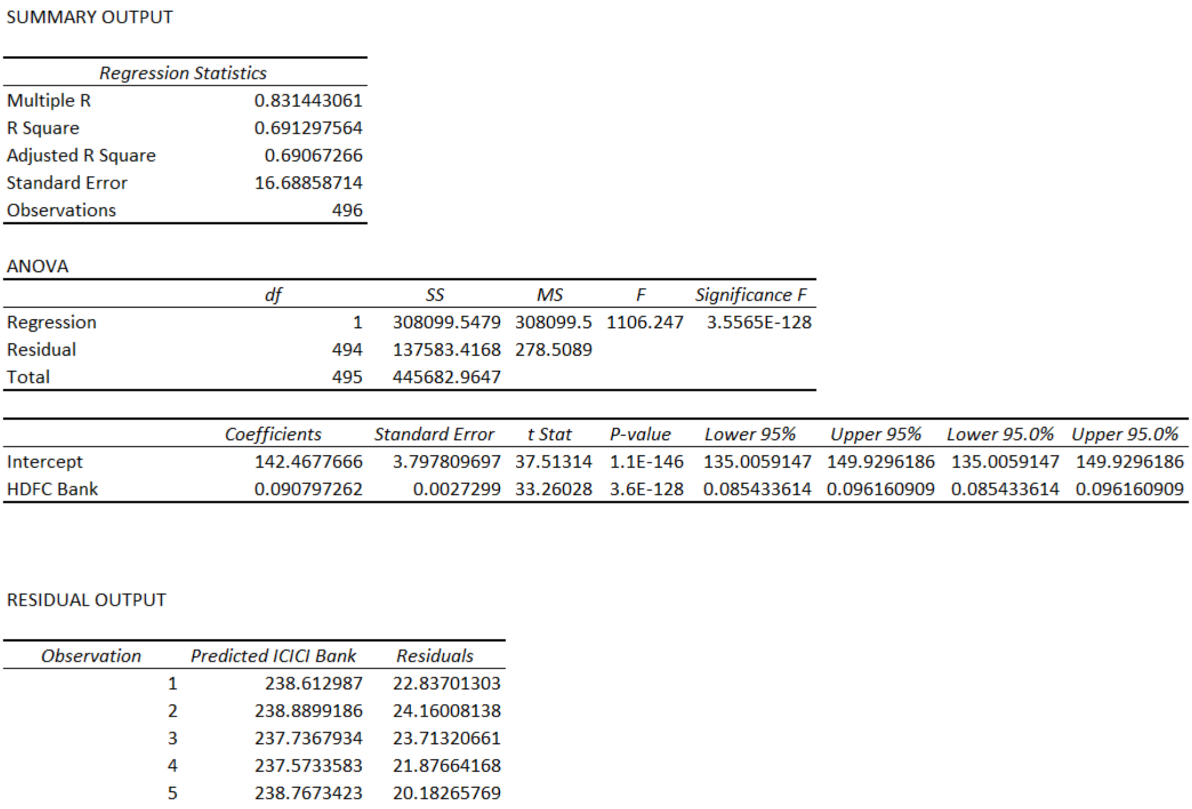
HDFC = Price of ICICI * 7.613 – 663.677

I'm assuming, you are familiar with the above equation. For those who are not familiar, I'd suggest you to read the previous two chapters. However here is the quick summary – the equation is trying to predict the price of HDFC using the price of ICICI.

Or in other words, we are trying to 'express' the price of HDFC in terms of ICICI.

Now, let us reverse this – I will set ICICI as dependent and HDFC as the independent.

Here is how the results look –



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	16.68858714
Observations	496

ANOVA

	df	SS	MS	F	Significance F
Regression	1	308099.5479	308099.5	1106.247	3.5565E-128
Residual	494	137583.4168	278.5089		
Total	495	445682.9647			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	142.4677666	3.797809697	37.51314	1.1E-146	135.0059147	149.9296186	135.0059147	149.9296186
HDFC Bank	0.090797262	0.0027299	33.26028	3.6E-128	0.085433614	0.096160909	0.085433614	0.096160909

RESIDUAL OUTPUT

Observation	Predicted ICICI Bank	Residuals
1	238.612987	22.83701303
2	238.8899186	24.16008138
3	237.7367934	23.71320661
4	237.5733583	21.87664168
5	238.7673423	20.18265769

The equation is –

ICICI = HDFC * 0.09 + 142.4677

So for the given two stocks, you can regress two ways by reordering which stock is dependent and which one is the independent variable.

However, the question is – how do you decide which one should be marked dependent and which one as independent. Or in other words, which order makes the most sense.

The answer to this depends on three things –

1. Standard Error
2. Standard Error of intercept
3. The ratio of the above two variables.

Remember, the linear equation above, essentially expresses the variation of price of ICICI in terms of HDFC (refer to the equation above). This expression or explanation of the price variation of one stock by keeping the price of the other stock as a reference can never be 100%. If it was 100%, then there is no play here at all.

Having said so, the equation should be strong enough to explain the variation in price of the dependent variable as much as possible, keeping the independent variable in perspective. The stronger this is, the better it is.

This leads us to the next obvious question – how do we figure out how strong the linear regression equation is? This is where the ratio –

Standard Error of Intercept / Standard Error comes into play. To understand this ratio, we need to understand both the numerator and the denominator before talking about the ratio itself.

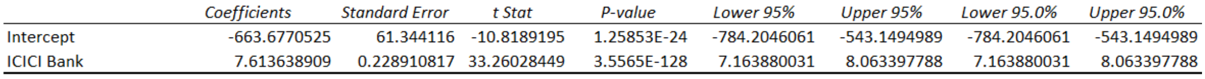
10.2 – Back to residuals

Here is the linear regression equation of ICICI as independent and HDFC as the dependent –

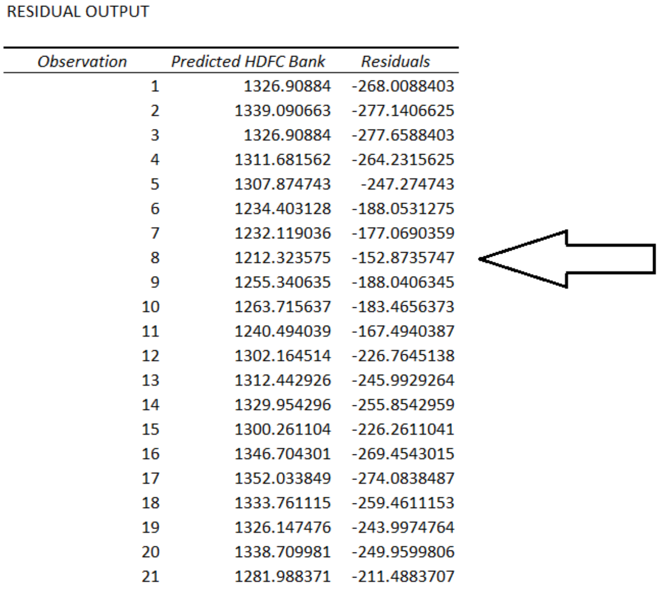
HDFC = Price of ICICI * 7.613 – 663.677

This essentially means, if I know the price of ICICI, I should be able to predict the price of HDFC. However, in reality, there is a difference between the predicted price of HDFC and the actual price. This difference is called the ‘Residuals’.

Here is the snapshot of the residuals when we try and explain the price of HDFC keeping ICICI as the independent variable –



	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-663.6770525	61.344116	-10.8189195	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.26028449	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788



RESIDUAL OUTPUT

Observation	*Predicted HDFC Bank*	*Residuals*
1	1326.90884	-268.0088403
2	1339.090663	-277.1406625
3	1326.90884	-277.6588403
4	1311.681562	-264.2315625
5	1307.874743	-247.274743
6	1234.403128	-188.0531275
7	1232.119036	-177.0690359
8	1212.323575	-152.8735747
9	1255.340635	-188.0406345
10	1263.715637	-183.4656373
11	1240.494039	-167.4940387
12	1302.164514	-226.7645138
13	1312.442926	-245.9929264
14	1329.954296	-255.8542959
15	1300.261104	-226.2611041
16	1346.704301	-269.4543015
17	1352.033849	-274.0838487
18	1333.761115	-259.4611153
19	1326.147476	-243.9974764
20	1338.709981	-249.9599806
21	1281.988371	-211.4883707

When I talk about the regression equation and the residuals, usually, I get one common question – what is the use of regression if there is a residual each and every time? Or in other words, how can we rely on an equation, which fails to predict accurately, even once.

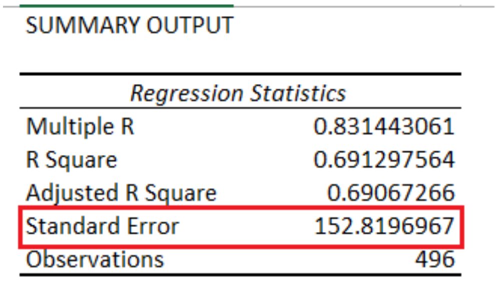
This is a fair question. If you look at the residuals above, they vary from a low of -288 to a high of 548, so using this equation to make any sort of prediction one price is futile.

But then, this was never about predicting the price of the dependent stock, given the price of an independent stock. It was always about the residuals!

Let me give you a heads-up here – the residuals display a certain behaviour. If we can understand this behaviour and figure a pattern within it, then we can rework backwards to construct a trade. This trade obviously involves buying and selling the two stocks simultaneously, hence this qualifies as a pair trade.

Over the next few chapter, we will dwell deeper into this. However, for now, let’s talk about the ‘Standard Error’, the denominator in the **Standard Error of Intercept / Standard Error** equation.

The standard error is one of the variables which gets reported when you run a linear regression operation. Here is the snapshot showing the same –

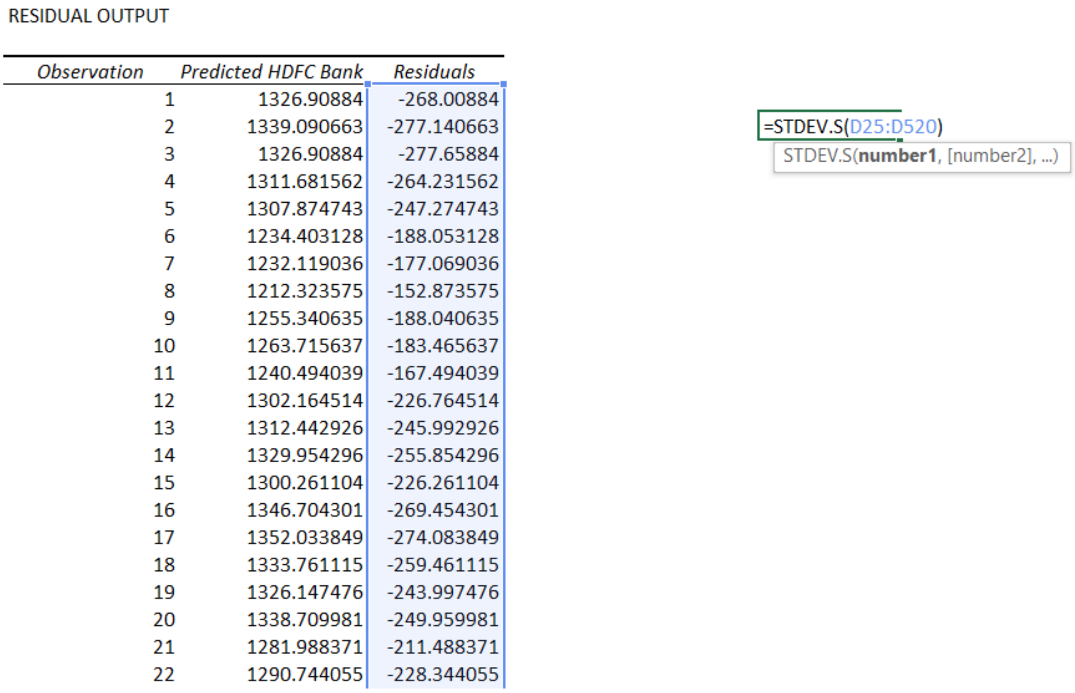


SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

The standard error is defined as the standard deviation of the residuals. Remember, the residuals itself is a time series array. So if you were to calculate the standard deviation of the residuals, then you get the standard error.

In fact, let me manually calculate the standard error of the residuals, I'm doing this for X = ICICI and y = HDFC



RESIDUAL OUTPUT

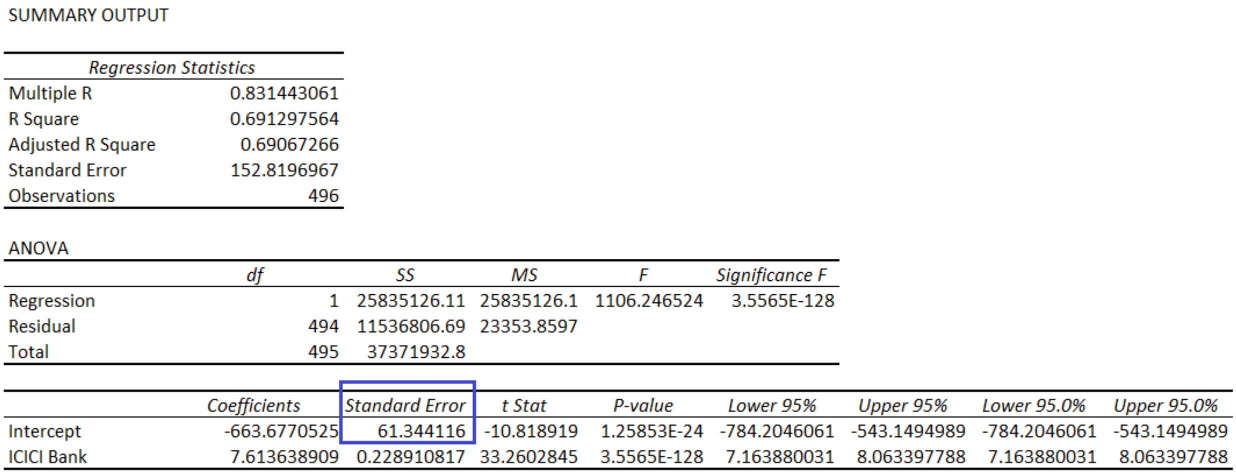
Observation	*Predicted HDFC Bank*	*Residuals*
1	1326.90884	-268.00884
2	1339.090663	-277.140663
3	1326.90884	-277.65884
4	1311.681562	-264.231562
5	1307.874743	-247.274743
6	1234.403128	-188.053128
7	1232.119036	-177.069036
8	1212.323575	-152.873575
9	1255.340635	-188.040635
10	1263.715637	-183.465637
11	1240.494039	-167.494039
12	1302.164514	-226.764514
13	1312.442926	-245.992926
14	1329.954296	-255.854296
15	1300.261104	-226.261104
16	1346.704301	-269.454301
17	1352.033849	-274.083849
18	1333.761115	-259.461115
19	1326.147476	-243.997476
20	1338.709981	-249.959981
21	1281.988371	-211.488371
22	1290.744055	-228.344055

=STDEV.S(D25:D520)

STDEV.S(**number1**, [number2], ...)

And excel tells me the standard deviation is **152.665**. The standard error as reported in the summary output is **152.819**. The minor difference can be ignored.

The 'Standard Error of the Intercept', is a little tricky. It does get reported in the regression report, and here is the standard error of the intercept with x = ICICI and y = HDFC



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	SS	MS	F	Significance F
Regression	1	25835126.11	25835126.1	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.8597		
Total	495	37371932.8			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-663.6770525	61.344116	-10.818919	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.2602845	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

Recall, the regression equation –

y=M*x+ C

Where,

M = Slope

C = Intercept

If you realize, here both M and C are estimates. And how are they estimated? They are estimated based on the historical data provided to the regression algorithm. The data can obviously contain noise components and few outliers. This implies that there is a scope for the estimates can go wrong.

The Standard Error of the Intercept is the measure of the variance of estimated intercept. It helps up understand by what degree the intercept itself can vary. So in a sense, this is somewhat similar to the 'Standard Error' itself. To summarize –

- Standard Error of Intercept – The variance of the intercept
- Standard Error – The variance of the residuals.

Now that we have defined both these variables, let's bring back the 'Error Ratio'. Please note, the term 'Error Ratio' is not a standard term, I've come up with it for ease of understanding.

Anyway, the error ratio, as we know –

Error Ratio = Standard Error of Intercept / Standard Error

I'm calculated the same for –

1. ICICI as X and HDFC as y = 0.401
2. HDFC as X and ICICI as y = 0.227

The decision to designate X and Y to stocks depends on the value of the error ratio. The lower the better. Since HDFC as X and ICICI as y offers the lowest error ratio, we will designate HDFC as the independent variable (X) and ICICI as the dependent variable (Y).

I'd love to explain the reason as to why we are using the error ratio as the key input for designating X and Y, but I guess I will hold back. I'll revisit this again when I take up pair trade example. For now, remember to calculate the error ratio and estimate which stock should be dependent and which one will be the independent.

You can download the excel sheet used in this chapter **here**.

Key takeaways from this chapter

1. X is the independent stock and Y is the dependent stock
2. The decision to figure out which stock is X and which one should be Y depends on 'Error Ratio'
3. Both the slope and the intercept from the linear regression equation are estimates
4. Error Ratio = Standard Error of the Intercept / Standard Error
5. Standard error is the standard deviation of the residuals
6. Standard error of intercept gives you a sense of the variance of the intercept
7. Regress Stock 1 with Stock 2 and also Stock 2 with Stock 1, whichever offers the lowest error ratio defines which stock is dependent and which one is independent
8. Residuals display certain properties, studying which can help identify pair trading pattern

CHAPTER 11

PTM2, C4 – The ADF test

11.1 – Co-Integration of two-time series

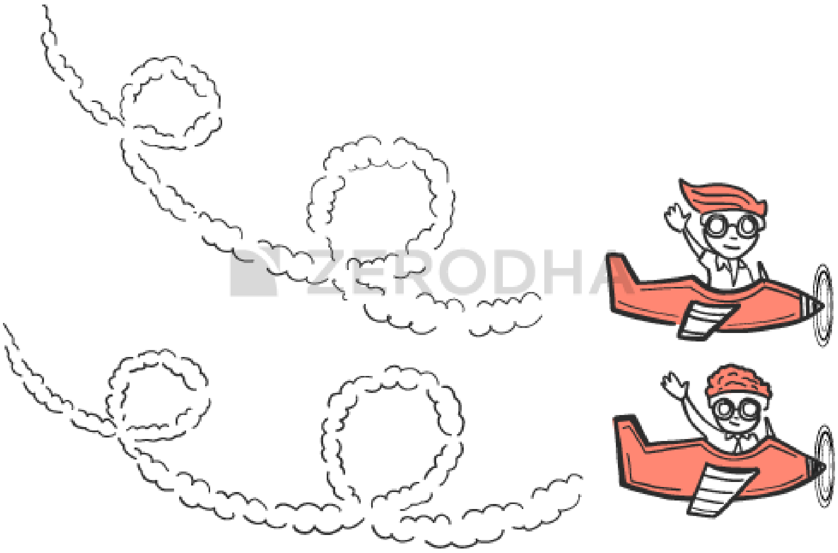
I guess this chapter will get a little complex. We would be skimming the surface of some higher order statistical theory. I will try my best and stick to practical stuff and avoid all the fluff. I'll try and explain these things from a trading point of view, but I'm afraid, some amount of theory will be necessary for you to know.

Given the path ahead I think it is necessary to re-rack our learnings so far and put some order to it. Hence let me just summarize our journey so far –

1. Starting from Chapter 1 to 7, we discussed a very basic version of a pair trade. We discussed this simply to lay out a strong foundation for the higher order pair trading technique, which is generally known as the relative value trade
2. The relative value trade requires the use of linear regression
3. In linear regression, we regress an independent variable, X against a dependent variable Y.
4. When we regress – some of the outputs that are of interest are the intercept, slope, residuals, standard error, and the standard error of the intercept
5. The decision to classify a stock as dependent and independent really depends on the error ratio.
6. We calculate the error ratio by interchanging both X and Y. The one which offers the lowest error ratio will define which stock is X and which on as Y.

I hope you have read and understood everything that we have discussed up to this point. If not, I'd suggest you read the chapters again, get clarity, and then proceed.

Recollect, in the previous chapter, we discussed the residuals. In fact, I also mentioned that the bulk of the focus going forward will be on the residuals. It is time we study the residuals in more detail and try and establish the kind of behaviour the residuals exhibit. In our attempt to do this, we will be introduced to two new jargons – Cointegration and Stationarity.



Generally speaking, if two-time series are 'co integrated' (stock X and stock Y in our case), then it means, that the two stocks move together and if at all there is a deviation from this movement, it is either temporary or can be attributed to a stray event, and one can expect the two-time series to revert to its regular orbit i.e. converge and move together again. Which is exactly what we want while pair trading. This means to say, the pair that we choose to pair trade on, should be cointegrated.

So the question is – how do we evaluate if the two stocks are cointegrated?

Well, to check if the two stock is cointegrated, we first need to run a linear regression on the two stocks, then take up the residuals obtained from the linear regression algorithm, and check if the residual is 'stationary'.

If the residuals are stationary, then it implies that the two stocks are cointegrated, if the two stocks are cointegrated, then the two stocks move together, and therefore the 'pair' is ripe for tracking pair trading opportunity.

Here is an interesting way to look at this – one can take any two-time series and apply regression, the regression algorithm will always throw out an output. How would one know if the output is reliable? This is where stationarity comes into play. The regression equation is valid if and only if residuals are stationary. If the residuals are not stationary, regression relation shouldn't be used.

Speculating and setting up trades on a co-integrated time series is a lot more meaningful and is independent of market direction.

So, essentially, this boils down to figuring out if the residuals are stationary or not.

At this point, I can straight away show you how to check if the residuals are stationary or not, there is a simple test called the 'ADF test' to do this – frankly, this is all you need to know. However, I think you are better off if you spend few minutes to understand what 'Stationarity' really means (without actually deep diving into the quants).

So, read the following section only if you are curious to know more, else go to the section which talks about ADF test.

11.2 Stationary and non-stationary series

A time series is considered 'Stationary' if it follows three 3 simple statistical conditions. If the time series partially satisfies these conditions, like 2 out of 3 or 1 out of 3, then the stationarity is considered weak. If none of the three conditions are satisfied, then the time series is 'non-stationary'.

The three simple statistical conditions are –

- The **mean** of the series should be same or within a tight range
- The **standard deviation** of the series should be within a range

- There should be no **autocorrelation** within the series – this means any particular value in the time series – say value ‘n’, should not be dependent on any other value before ‘n’. Will talk more about this at a later stage.

While pair trading, we only look for pairs which exhibit complete stationarity. Non-stationary series or weak stationary series will not work for us.

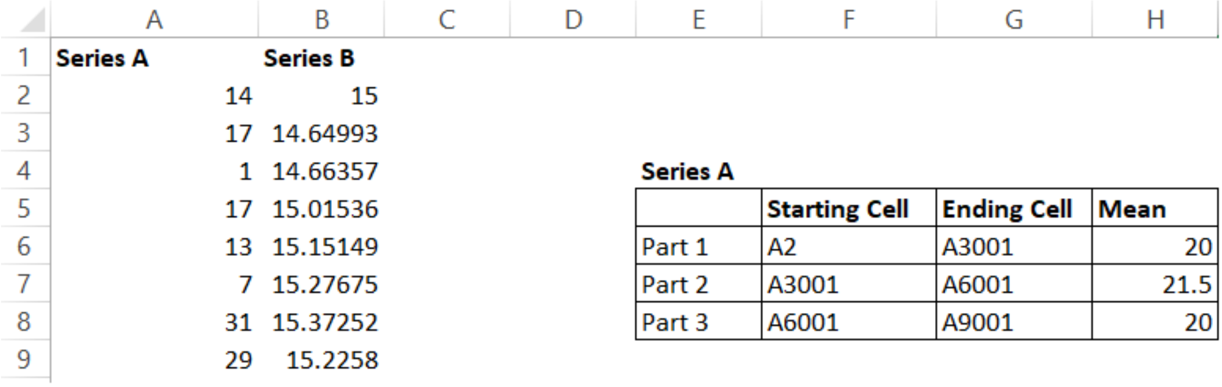
I guess it is best to take up an example (like a sample time series) and figure out what the above three conditions really mean and hopefully, that will help you understand ‘stationarity’ better.

For the sake of this example, I have two-time series data, with 9000 data points in each. I’ve named them Series A and Series B, and on this time series data, I will evaluate the above three stationarity conditions.

Condition 1 – The mean of the series should be same or within a tight range

To evaluate this, I will split each of the time series data into 3 parts and calculate the respective mean for each part. The mean for all three different parts should be around the same value. If this is true, then I can conclude that the mean will more or less be the same even when new data points flow in the future.

So let us go ahead and do this. To begin with, I’m splitting the Series A data into three parts and calculating its respective means, here is how it looks –



	A	B
1	**Series A**	**Series B**
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258

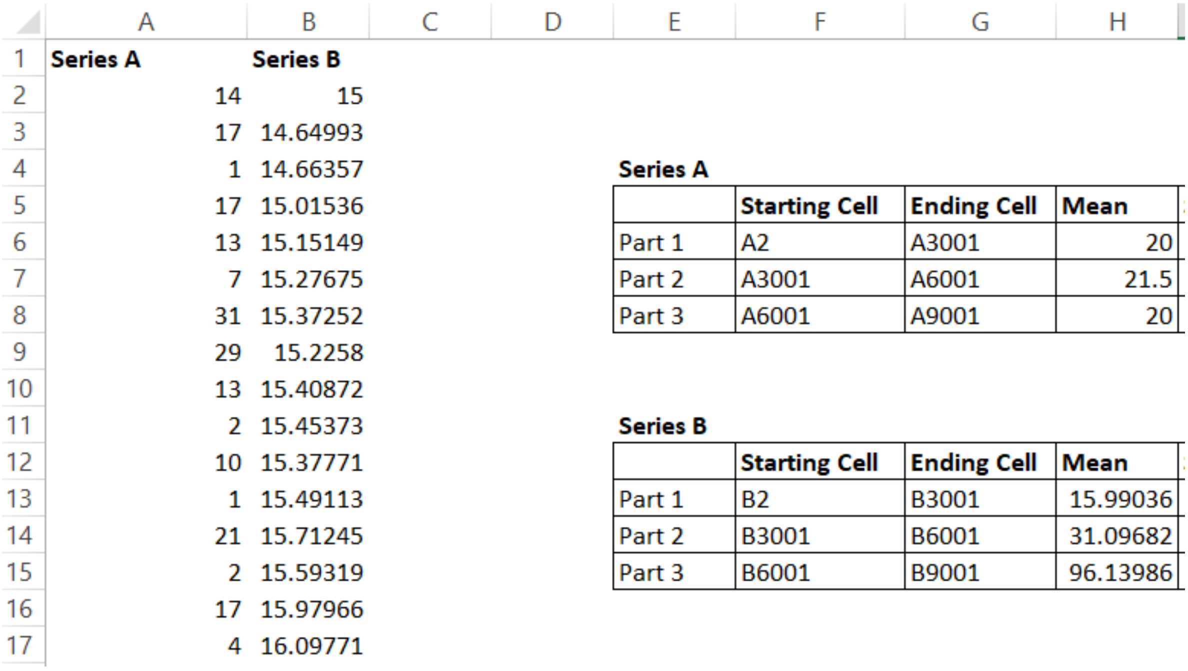
Series A

	Starting Cell	**Ending Cell**	**Mean**
Part 1	A2	A3001	20
Part 2	A3001	A6001	21.5
Part 3	A6001	A9001	20

Like I mentioned, I have 9000 data points in Series A and Series B. I have split Series A data points into 3 parts and as you can see, I've even highlighted the starting and ending cells for these parts.

The mean for all the three parts are similar, clearly satisfying the first condition.

I've done the same thing for Series B, here is how the mean looks –



	A	B
1	Series A	Series B
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258
10	13	15.40872
11	2	15.45373
12	10	15.37771
13	1	15.49113
14	21	15.71245
15	2	15.59319
16	17	15.97966
17	4	16.09771

Series A

	Starting Cell	Ending Cell	Mean
Part 1	A2	A3001	20
Part 2	A3001	A6001	21.5
Part 3	A6001	A9001	20

Series B

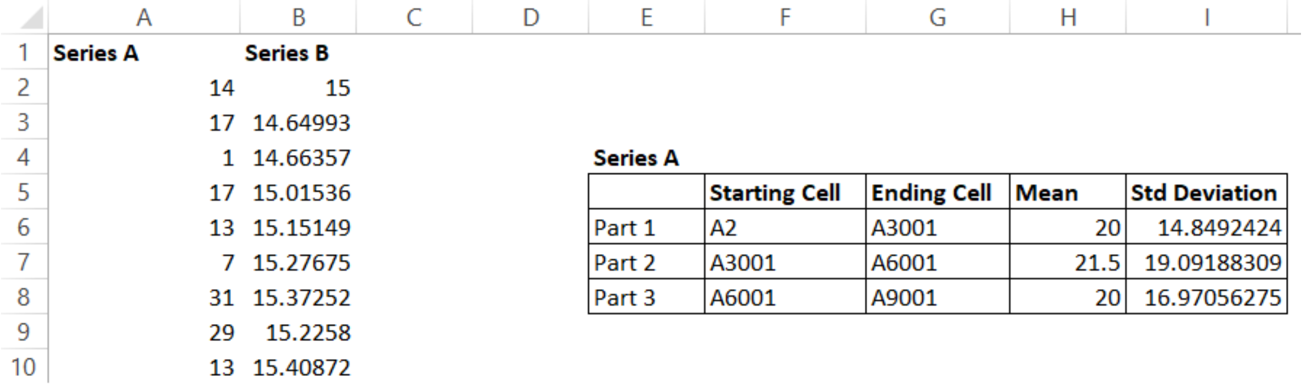
	Starting Cell	Ending Cell	Mean
Part 1	B2	B3001	15.99036
Part 2	B3001	B6001	31.09682
Part 3	B6001	B9001	96.13986

Now as you can see, the mean for Series B swings quite wildly and thereby not satisfying the first condition for stationarity.

Condition 2 -The standard deviation should be within a range.

I'm following the same approach here – I will go ahead and calculate the standard deviation for all the three parts for both the series and observe the values.

Here is the result obtained for Series A –



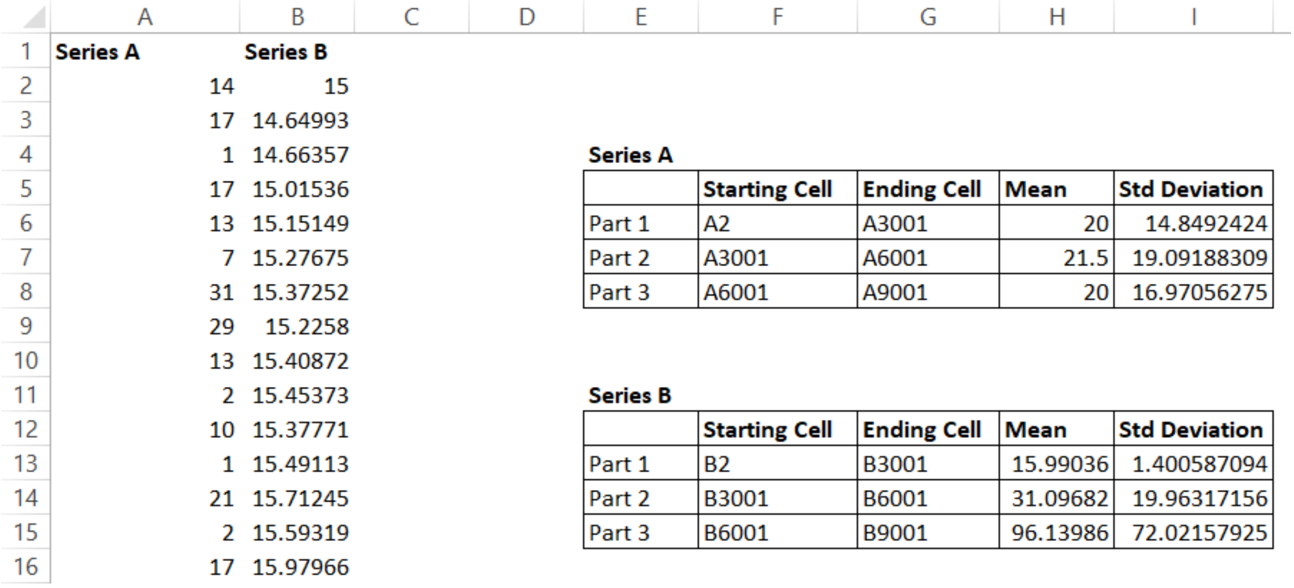
	A	B
1	Series A	Series B
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258
10	13	15.40872

Series A

	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	A2	A3001	20	14.8492424
Part 2	A3001	A6001	21.5	19.09188309
Part 3	A6001	A9001	20	16.97056275

The standard deviation oscillates between 14-19%, which is quite 'tight' and therefore qualifies the 2ndstationarity condition.

Here is how the standard deviation works out for Series B –



	A	B
1	Series A	Series B
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258
10	13	15.40872
11	2	15.45373
12	10	15.37771
13	1	15.49113
14	21	15.71245
15	2	15.59319
16	17	15.97966

Series A

	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	A2	A3001	20	14.8492424
Part 2	A3001	A6001	21.5	19.09188309
Part 3	A6001	A9001	20	16.97056275

Series B

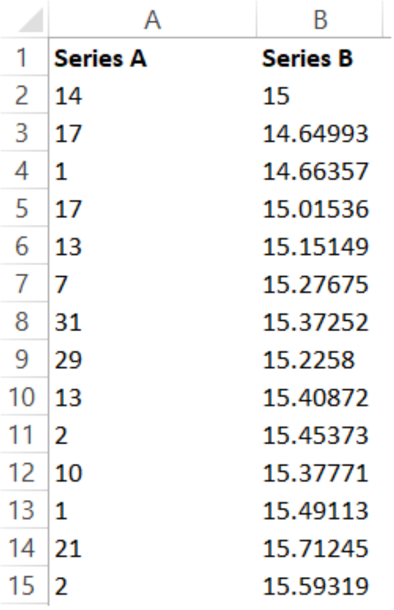
	Starting Cell	Ending Cell	Mean	Std Deviation
Part 1	B2	B3001	15.99036	1.400587094
Part 2	B3001	B6001	31.09682	19.96317156
Part 3	B6001	B9001	96.13986	72.02157925

Notice the difference? The range of standard deviation for Series B is quite random. Series B is clearly not a stationary series. However, Series A looks stationary at this point. However, we still need to evaluate the last condition i.e. the autocorrelation bit, let us go ahead and do that.

Condition 3 – There should be no autocorrelation within the series

In layman words, autocorrelation is a phenomenon where any value in the time series is not really dependent on any other value before it.

For example, have a look at the snapshot below –



	A	B
1	**Series A**	**Series B**
2	14	15
3	17	14.64993
4	1	14.66357
5	17	15.01536
6	13	15.15149
7	7	15.27675
8	31	15.37252
9	29	15.2258
10	13	15.40872
11	2	15.45373
12	10	15.37771
13	1	15.49113
14	21	15.71245
15	2	15.59319

The 9th value in Series A is 29, and if there is no autocorrelation in this series, the value 29 is not really dependent on any values before it i.e. the values from cell 2 to cell 8.

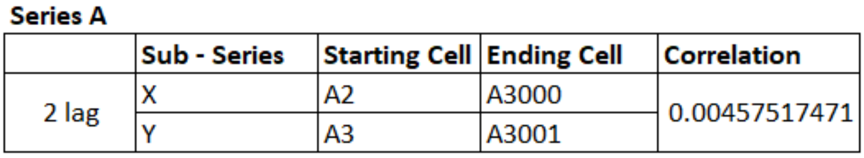
But the question is how do we establish this?

Well, there is a technique for this.

Assume there are 10 data points, I take the data from Cell 1 to Cell 9, call this series X, now take the data from Cell 2 to Cell 10, call this Series Y. Now, calculate the correlation between Series X and Y. This is called 1-lag correlation. The correlation should be near to 0.

I can do this for 2 lag as well – i.e. between Cell 1 to Cell 8, and then between Cell 3 to Cell 10, again, the correlation should be close to 0. If this is true, then it is safe to assume assumed that the series is not auto correlated, and hence the 3rd condition for stationarity is proved.

I've calculated 2 lag correlation for Series A, and here is how it looks –

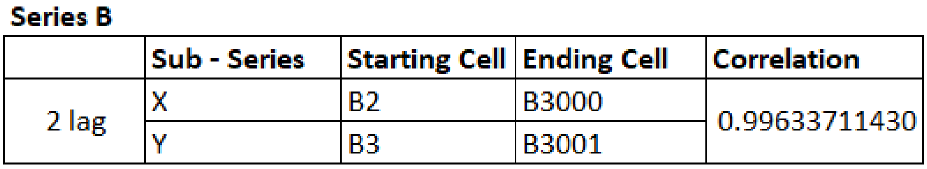


Series A

	Sub - Series	Starting Cell	Ending Cell	Correlation
2 lag	X	A2	A3000	0.00457517471
	Y	A3	A3001	

Remember, I'm subdividing Series A into two parts and creating two subseries i.e. series X and series Y. The correlation is calculated on these two subseries. Clearly, the correlation is close to zero and with this, we can safely conclude that Time Series A is stationary.

Let's do this for Series B as well.



Series B

	Sub - Series	Starting Cell	Ending Cell	Correlation
2 lag	X	B2	B3000	0.99633711430
	Y	B3	B3001	

I've taken a similar approach, and the correlation as you can see is quite close to 1.

So, as you can see all the conditions for stationarity is met for Series A – which means the time series is stationary. While Series B is not.

I know that I've taken a rather unconventional approach to explaining stationarity and co-integration. After all, no statistical explanation is complete without those scary looking formulas. But this is a deliberate approach and I thought this would be the best possible way to discuss these topics, as eventually, our goal is to learn how to pair trade efficiently and not really deep dive into statistics.

Anyway, you could be thinking if it is really required for you to do all of the above to figure out if the time series (residuals) are indeed stationary. Well, like I said before, this is not required.

We only need to look at the results of something called as the 'The ADF Test', to establish if the time series is stationary or not.

11.3 – The ADF test

The augmented Dickey-Fuller or the ADF test is perhaps one of the best techniques to test for the stationarity of a time series. Remember, in our case, the time series in consideration is the residuals series.

Basically, the ADF test does everything that we discussed above, including a multiple lag process to check the autocorrelation within the series. Here is something you need to know – the output of the ADF test is not a definitive 'Yes – this is a stationary series' or 'No – this is not a stationary series'. Rather, the output of the ADF test is a probability. It tells us the probability of the series, not being stationary.

For example, if the output of the ADF test a time series is 0.25, then this means the series has a 25% chance of not being stationary or in other words, there is a 75% chance of the series being stationary. This probability number is also called 'The P value'.

To consider a time series stationary, the P value should be as low as 0.05 (5%) or lower. This essentially means the probability of the time series is stationary is as high as 95% (or higher).

Alright, so how do you run an ADF test?

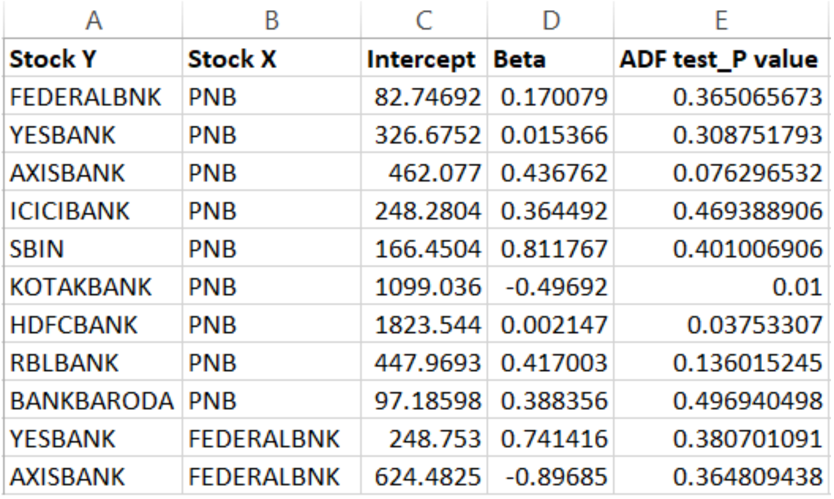
Frankly, this is a highly complex process and unfortunately, I could not find a single source online which will help you run an ADF test for free. I do have an excel sheet (which has a paid plugin) to run an ADF test, but unfortunately, I cannot share it here. If I could, I would have.

If you are a programmer, I've been told that there are Python plugins easily available to run an ADF test, so you could try that.

But if you are a non-programmer like me, then you will be stuck at this stage. So here is what I will do, once in a weak or 15 days, I will try and upload a 'Pair Data' sheet, which will contain the following information of the best possible combination of pairs, this includes –

1. You will know which stock is X and which stock is Y
2. You will know the intercept and Beta of this combination
3. You will also know the p-value of the combination

The look back period for generating this is 200 trading days. I've restricted this just to banking stocks, but hopefully, I can include more sectors going forward. To help you understand this better, here is the snapshot of the latest Pair Datasheet for banking stocks –



A	B	C	D	E
Stock Y	**Stock X**	**Intercept**	**Beta**	**ADF test_P value**
FEDERALBNK	PNB	82.74692	0.170079	0.365065673
YESBANK	PNB	326.6752	0.015366	0.308751793
AXISBANK	PNB	462.077	0.436762	0.076296532
ICICIBANK	PNB	248.2804	0.364492	0.469388906
SBIN	PNB	166.4504	0.811767	0.401006906
KOTAKBANK	PNB	1099.036	-0.49692	0.01
HDFCBANK	PNB	1823.544	0.002147	0.03753307
RBLBANK	PNB	447.9693	0.417003	0.136015245
BANKBARODA	PNB	97.18598	0.388356	0.496940498
YESBANK	FEDERALBNK	248.753	0.741416	0.380701091
AXISBANK	FEDERALBNK	624.4825	-0.89685	0.364809438

The first line suggests that Federal Bank as Y and PNB as X is a viable pair. This also means, that the regression of Federal as Y and PNB as X and Federal as X and PNB as Y was conducted and the error ratio for both the combination was calculated, and it was found that Federal as Y and PNB as X had the least error ratio.

Once the order has been figured out (as in which one is Y and which one is X), the intercept and Beta for the combination has also been calculated. Finally, the ADF was conducted and the P value was calculated. If you see, the P value for Federal Bank as Y and PNB as X is 0.365.

In other words, this is not a combination you should be dealing with as the probability of the residuals being stationary is only 63.5%.

In fact, if you look at the snapshot above, you will find only 2 pairs which have the desired p-value i.e. Kotak and PNB with a P value of 0.01 and HDFC and PNB with a P value of 0.037.

The p values don't usually change overnight. Hence, for this reason, I check for p-value once in 15 or 20 days and try and update them here.

I think we have learned quite a bit in this chapter. A lot of information discussed here could be new for most of the readers. For this reason, I will summarize all the things you should know about Pair trading at this point –

1. The basic premise of pair trading
2. Basic overview of linear regression and how to perform one
3. In linear regression, we regress an independent variable, X against a dependent variable Y.
4. When we regress – some of the outputs that are of interest are the intercept, slope, residuals, standard error, and the standard error of the intercept
5. The decision to classify a stock as dependent and independent really depends on the error ratio.
6. We calculate the error ratio by interchanging both X and Y. The one which offers the lowest error ratio will define which stock is X and which on as Y
7. The residuals obtained from the regression should be stationary. If they are stationary, then we can conclude that the two stocks are co-integrated
8. If the stocks are cointegrated, then they move together
9. Stationarity of a series can be evaluated by running an ADF test.

If you are not clear on any of the points above, then I'd suggest you give this another shot and start reading from Chapter 7.

In the next chapter, we will try and take up an example of a pair trade and understand its dynamics.

You can **download the Pair Data** sheet, updated on 11th April 2018.

Lastly, this module (and this chapter, in particular) could not have been possible without the inputs from my good friend and an old partner, **Prakash Lekkala**. So I guess, we all need to thank him

---

Key takeaways from this chapter –

1. If two stocks move together, then they are also cointegrated
2. You can pair trade on stocks which are cointegrated
3. If the residuals obtained from linear regression is stationary, then it implies the two stocks are co-integrated
4. A time series is considered stationary if the series has a constant mean, constant standard deviation, and no autocorrelation
5. The check for stationarity can be done by an ADF test
6. The p-value of the ADF test should be 0.05% or lower for the series to be considered stationary.

CHAPTER 12

Trade Identification

12.1 – Trading the equation

At this stage, we have discussed pretty much all the background information we need to know about Pair trading. We now have to patch things together and understand how all these concepts make sense while taking up a pair trade.

Let's start with the basic equation again. I understand we have gone through this equation earlier in this module, but I want you to relook at this equation from a trader's perspective. I want you to think about ways in which you can trade this equation. I want you to see opportunities here. This is where everything starts to culminate.

y = M*x + c

What is this equation essentially trying to tell you? Well, frankly, it depends on how your perspective of this equation. You can look at it from two different perspectives –

1. As a statistician
2. As a trader

Since we are dealing with two stocks here, the **statistician** would look at this as an equation where the stock price of a dependent stock 'y' is being explained with respect to an independent stock price 'x'. This process of 'price explanation' generates two other variables i.e. the slope (or beta) 'M' and the intercept 'c'.

So in an ideal world, the stock price of y should be exactly equal to the Beta times X plus the intercept.

But we know that this is not true, there is always a variation in this equation which leads to the difference between the actual stock price of Y and the predicted stock price of Y. This difference is also termed as the 'residual' or the error term.

In fact, we can extend the above equation to include the residuals and with that, the equation would look like this –

$$y = M*x + c + \varepsilon$$

Where, ε represents the error or the residual of the equation. Of course, by now we are even familiar with the stationarity of the residuals which adds more sanctity to the above equation.

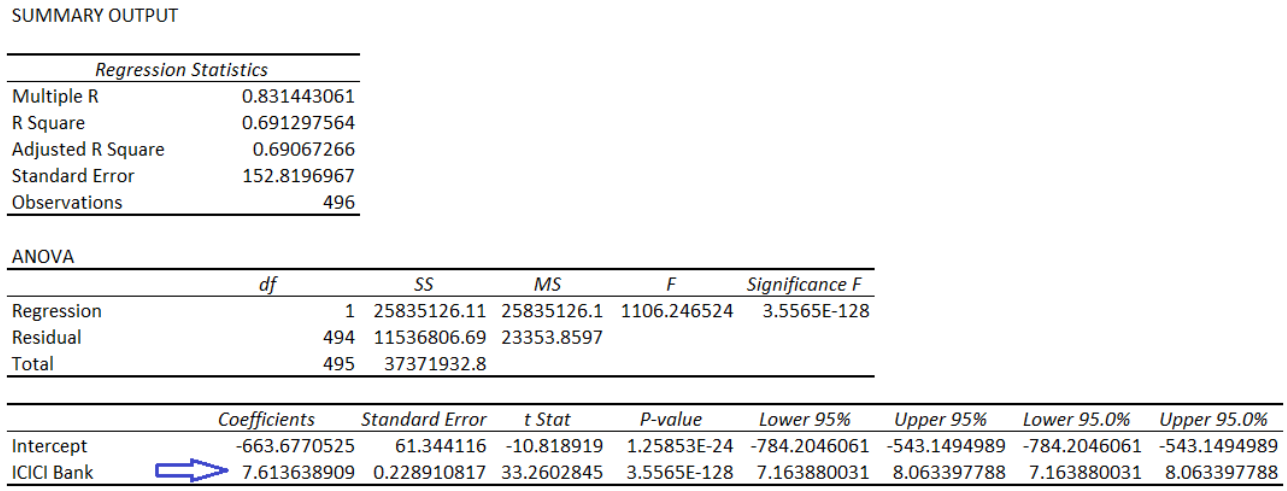
Fair enough, now for the interesting bit – how would a **trader** look at this equation? Let me repost the equation again –

$$y = M*x + c + \varepsilon$$

Let us break this equation into smaller pieces –

y = M*x, this essentially means, the price of the dependent stock ‘y’ is equal to the independent stock price ‘x’, multiplied by the slope M. Well, the slope is essentially the beta and it tells us how many stocks of x would equal the price of y.

For example, here is the linear regression output of HDFC Bank (y) vs ICICI Bank (x) –



SUMMARY OUTPUT

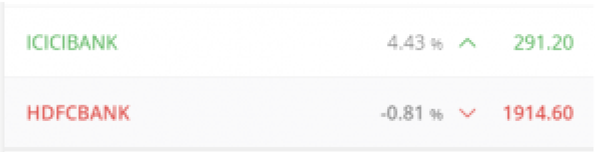
Regression Statistics	
Multiple R	0.831443061
R Square	0.691297564
Adjusted R Square	0.69067266
Standard Error	152.8196967
Observations	496

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	25835126.11	25835126.1	1106.246524	3.5565E-128
Residual	494	11536806.69	23353.8597		
Total	495	37371932.8			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-663.6770525	61.344116	-10.818919	1.25853E-24	-784.2046061	-543.1494989	-784.2046061	-543.1494989
ICICI Bank	7.613638909	0.228910817	33.2602845	3.5565E-128	7.163880031	8.063397788	7.163880031	8.063397788

And here is the snapshot of the prices of ICICI and HDFC –



ICICIBANK	4.43 % ^	291.20
HDFCBANK	-0.81 % ∨	1914.60

Now, this means, the price of HDFC Bank is roughly equal to the price of ICICI times the Beta. So, 1914 = 291 *7.61.

Don’t jump in to do the math, I know that does not add up

But for a moment, assume if this equation were to be true, then, in other words, this essentially means 7.61 shares of ICICI equals 1 share of HDFC. This is an important conclusion.

This also means, if I were to go long on one share of HDFC and short on 7.61 shares of ICIC, then I’m essentially long and short at the same time, hence I’ve hedged away a large amount of directional risk. Don’t forget the basic premise here, we are considering these two stocks because they are co-integrated in the first place.

So here is the equation again –

$$y = M*x + c + \varepsilon$$

If this equation were to be true, then by going long and short on y and x, we are hedging away the directional risk associated with this pair.

This leaves us with the 2nd part of the equation i.e. **$c + \varepsilon$**

As you know, C is the intercept. Now, at this point, I want you to recollect the 'Error Ratio' which we discussed in chapter 10.

Error Ratio = Standard Error of Intercept / Standard Error.

As you may recollect, we discussed the lower the error ratio, the better it is. Mathematically, this also implies that we are looking at pairs which have a low intercept.

Again this is a very crucial point for you to note, we are selecting the pairs, such that the standard error of the intercept is low.

Remember, in this equation **$y = M*x + c + \varepsilon$** we are trying to establish a trade (or hedge) every element. We are hedging y with Mx. We are trying to minimize c or the intercept because we are not trading or hedging it. Therefore, the lower it is, the better for us.

This leaves us with just the residual or the ε.

Remember, the residual is a time series. We have even validated the stationarity of this series. Now, because the residual is a stationary time series, the properties of normal distribution can be quite beautifully applied. This means, I only need to track the residuals and trigger a trade when it hits the upper or lower standard deviation!

Generally speaking, a trade is initiated when –

1. Long on the pair (buy y, sell x) when the residuals hit -2 standard deviation (-2SD)
2. Short on the pair (sell y, buy x) when the residuals hit +2 standard deviations (+2SD)

Like in the first method, the idea here is to initiate a trade at the 2nd standard deviation and hold the trade till the residual reverts to mean. The SL can be kept at 3SD for both the trades. More on this in the next chapter.

I know this is a short chapter, but I will conclude it here, as I don't want to clutter your mind with other information.

It is important for you to understand this equation from a trader's perspective and figure out what exactly you are trading. Remember, we are only trading the residuals here. We are hedging away the stock price of y with x. The intercept is kept low, and the residual is traded.

Why is the residual tradable? Because its stationary and therefore, its behaviour is kind of predictable. In the next chapter, I'll try and take up a live trade and deal with the practical aspects of pair trading.

Key takeaways from this chapter

1. The pair trading equation is actually the main equation which we trade
2. Every element of the equation is looked into
3. We hedge the stock price of y with the stock price of x. The beta of x tells us the number stocks required to hedge 1 stock of y
4. By looking into the error ratio, we are ensuring the intercept is kept low. Please remember we are not hedging the intercept, hence this needs to be kept low
5. The residual is what we trade as it is stationary and follows the normal distribution quite well
6. A long trade is initiated when residuals hit -2SD. Likewise, a short trade is initiated when the residuals hit +2SD
7. Long on a pair requires us to go long on Y and short on X
8. Short on a pair requires us to go short on Y and long on X

9. When we initiate a pair trade, we expect the residual to hit the mean, so we hold until then
10. The SL can be kept at 3SD for both long and short trades

CHAPTER 13

Live Example -1

13.1 – Tracking the pair data

We have finally reached a point where we are through with all the background theory knowledge required for Pair Trading. I know most of you have been waiting for this

moment

In this last and final chapter of pair trading, we will take up an example of a live trade and discuss factors that influence the trade.

Here is a quick recap of pre-trade theory –

1. Basic overview of linear regression and how to perform one
2. Linear regression requires you to regress an independent variable X against a dependent variable Y

3. The output of linear regression includes the intercept, slope, residuals, standard error, and the standard error of the intercept
4. The decision to classify a stock as dependent (Y) and independent (X) depends the error ratio
5. Error ratio is defined as the ratio of standard error of intercept/standard error
6. We calculate the error ratio by interchanging both X and Y. The combination which offers the lowest error ratio will define which stock is assigned X and which on as Y
7. The residuals obtained from the regression should be stationary. If they are stationary, then we can conclude that the two stocks are co-integrated
8. If the stocks are cointegrated, then they move together
9. Stationarity of a series can be evaluated by running an ADF test
10. The ADF value of an ideal pair should be less than 0.05

Over the last few chapters, we have discussed each point in great details. These points help us understand which pairs are worth considering for pair trading. In a nutshell, we take any two stocks (from the same sector), run a linear regression on it, check the error ratio and identify which stock is X and which is Y. We now run an ADF test on the residual of the pair. A pair is considered worth tracking (and trading) only if the ADF is 0.05 or lower. If the pair qualifies this, we then track the residuals on a daily basis and try to spot trading opportunities.

A pair trade opportunity arises when –

1. The residuals hit -2 standard deviations (-2SD). This is a long signal on the pair, so we buy Y and sell X
2. The residuals hit +2 standard deviations (+2SD). This is a short signal on the pair, so we sell Y and buy X

Having said so, I generally prefer to initiate the trade when the residuals hit 2.5 SD or thereabouts. Once the trade is initiated, the stop loss is -3 SD for long trades and +3SD for short trades and the target is -1 SD and +1 SD for long and short trades respectively. This also means, once you initiate a pair trade, you will have to track the residual value to know where it lies and plan your trades. Of course, we will discuss more on this later in this chapter.

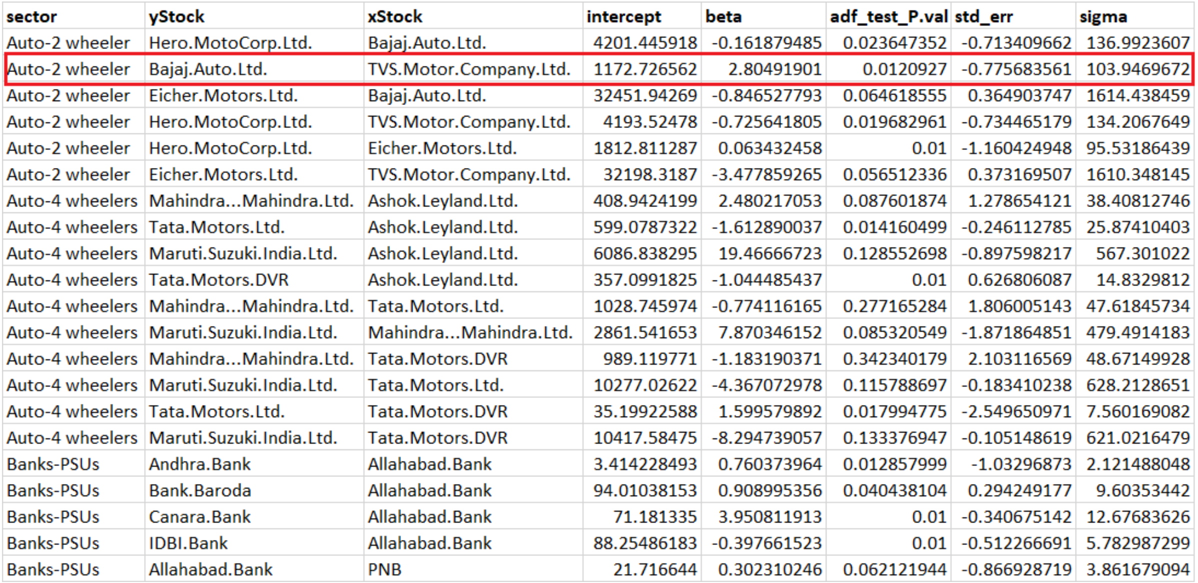
13.2 – Note for the programmers

In **Chapter 11**, I introduced the 'Pair Data' sheet. This sheet is an output of the Pair Trading Algo. The pair trading algo basically does the following –

1. Downloads the last 200-day closing prices of the underlying. You can do this from NSE's bhavcopy, in fact, automate the same by running a script.
2. The list of stock and its sector classification is already done. Hence the download is more organized
3. Runs a series of regressions and calculates the 'error ratio' for each regression. For example, if we are talking about RBL Bank and Kotak Bank, then the regression module would regress RBL (X) and Kotak (Y) and Kotak (X) and RBL (Y). The combination which has the lowest error ratio is considered and the other combination is ignored
4. The ADF test is applied on the residuals, for the combination which has the lowest error ratio.
5. A report (pair data) is generated with all the viable X-Y combination and its respective intercepts, beta, ADF value, standard error, and sigma are noted. I know we have not discussed sigma yet, I will shortly.

If you are a programmer, I would suggest you use this as a guideline to develop your own pair trading algo.

Anyway, in Chapter 11, I had briefly explained how to read the data from the Pair data, but I guess it's time to dig into the details of this output sheet. Here is the snapshot of the Pair data excel sheet –



sector	yStock	xStock	intercept	beta	adf_test_P.val	std_err	sigma
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Bajaj.Auto.Ltd.	4201.445918	-0.161879485	0.023647352	-0.713409662	136.9923607
Auto-2 wheeler	Bajaj.Auto.Ltd.	TVS.Motor.Company.Ltd.	1172.726562	2.80491901	0.0120927	-0.775683561	103.9469672
Auto-2 wheeler	Eicher.Motors.Ltd.	Bajaj.Auto.Ltd.	32451.94269	-0.846527793	0.064618555	0.364903747	1614.438459
Auto-2 wheeler	Hero.MotoCorp.Ltd.	TVS.Motor.Company.Ltd.	4193.52478	-0.725641805	0.019682961	-0.734465179	134.2067649
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Eicher.Motors.Ltd.	1812.811287	0.063432458	0.01	-1.160424948	95.53186439
Auto-2 wheeler	Eicher.Motors.Ltd.	TVS.Motor.Company.Ltd.	32198.3187	-3.477859265	0.056512336	0.373169507	1610.348145
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Ashok.Leyland.Ltd.	408.9424199	2.480217053	0.087601874	1.278654121	38.40812746
Auto-4 wheelers	Tata.Motors.Ltd.	Ashok.Leyland.Ltd.	599.0787322	-1.612890037	0.014160499	-0.246112785	25.87410403
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Ashok.Leyland.Ltd.	6086.838295	19.46666723	0.128552698	-0.897598217	567.301022
Auto-4 wheelers	Tata.Motors.DVR	Ashok.Leyland.Ltd.	357.0991825	-1.044485437	0.01	0.626806087	14.8329812
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.Ltd.	1028.745974	-0.774116165	0.277165284	1.806005143	47.61845734
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Mahindra...Mahindra.Ltd.	2861.541653	7.870346152	0.085320549	-1.871864851	479.4914183
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.DVR	989.119771	-1.183190371	0.342340179	2.103116569	48.67149928
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.Ltd.	10277.02622	-4.367072978	0.115788697	-0.183410238	628.2128651
Auto-4 wheelers	Tata.Motors.Ltd.	Tata.Motors.DVR	35.19922588	1.599579892	0.017994775	-2.549650971	7.560169082
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.DVR	10417.58475	-8.294739057	0.133376947	-0.105148619	621.0216479
Banks-PSUs	Andhra.Bank	Allahabad.Bank	3.414228493	0.760373964	0.012857999	-1.03296873	2.121488048
Banks-PSUs	Bank.Baroda	Allahabad.Bank	94.01038153	0.908995356	0.040438104	0.294249177	9.60353442
Banks-PSUs	Canara.Bank	Allahabad.Bank	71.181335	3.950811913	0.01	-0.340675142	12.67683626
Banks-PSUs	IDBI.Bank	Allahabad.Bank	88.25486183	-0.397661523	0.01	-0.512266691	5.782987299
Banks-PSUs	Allahabad.Bank	PNB	21.716644	0.302310246	0.062121944	-0.866928719	3.861679094

Look at the highlighted data. The Y stock is Bajaj Auto and X stock is TVS. Now because this combination is present in the report, it implies – Bajaj as Y and TVS as X has a lower standard error ratio, which further implies that Bajaj as X and TVS as Y is not a viable pair owing to higher error ratio, hence you will not find this combination (Bajaj as X and TVS as Y) in this report.

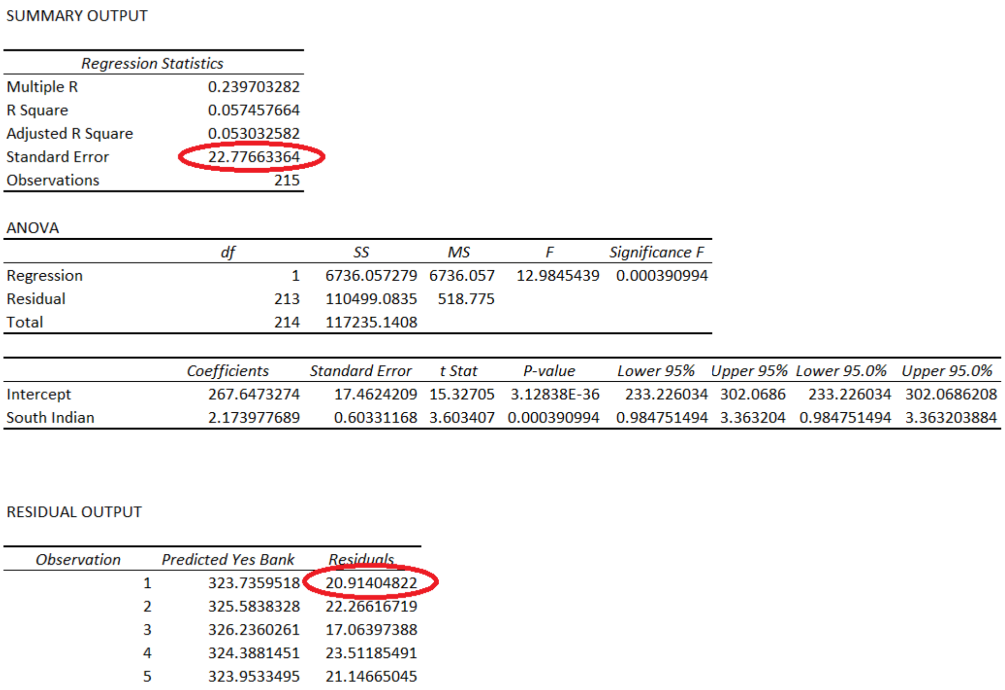
Along with identifying which one is X and Y, the report also gives you the following information –

1. Intercept – 1172.72
2. Beta – 2.804
3. ADF value – 0.012
4. Std_err – -0.77
5. Sigma – 103.94

I'm assuming (and hopeful) you are aware of the first three variables i.e. intercept, Beta, and ADF value so I won't get into explaining this all over again. I'd like to quickly talk about the last two variables.

Standard Error (or Std_err) as mentioned in the report is essentially a ratio of Today's residual over the standard error of the residual. Please note, this can get a little confusing here because there are two standard errors' we are talking about. The 2nd standard error is the standard error of the residual, which is reported in the regression output. Let me explain this with an example.

Have a look at the snapshot below –



SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.239703282
R Square	0.057457664
Adjusted R Square	0.053032582
Standard Error	22.77663364
Observations	215

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	6736.057279	6736.057	12.9845439	0.000390994
Residual	213	110499.0835	518.775		
Total	214	117235.1408			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	267.6473274	17.4624209	15.32705	3.12838E-36	233.226034	302.0686	233.226034	302.0686208
South Indian	2.173977689	0.60331168	3.603407	0.000390994	0.984751494	3.363204	0.984751494	3.363203884

RESIDUAL OUTPUT

Observation	*Predicted Yes Bank*	*Residuals*
1	323.7359518	20.91404822
2	325.5838328	22.26616719
3	326.2360261	17.06397388
4	324.3881451	23.51185491
5	323.9533495	21.14665045

This is the regression output summary of Yes Bank versus South Indian Bank. I've highlighted standard error (22.776). This is the standard error of the residuals. Do recollect, we have discussed this earlier in this module.

The second highlight is 20.914, which is the residual.

The std_err in the report is simply a ratio of –

Today's residual / Standard Error of the residual

= 20.92404/22.776

= 0.91822

Yes, I agree calling this number std_err is not the best choice, but please bear with it for

now

This number gives me information of how today's residual is position in the context of the standard distribution. This is the number which is the key trigger for the trade. A long position is hit if this number is -2.5 or higher with -3.0 as stop loss. A short position is initiated if this number reads +2.5 or higher with a stop loss at +3.0. In case of long, target is at -1 or lower and in case of short, the target is +1 or lower.

This also means, the std_err number has to be calculated on a daily basis and tracked to identify trading opportunities. More on this in a bit.

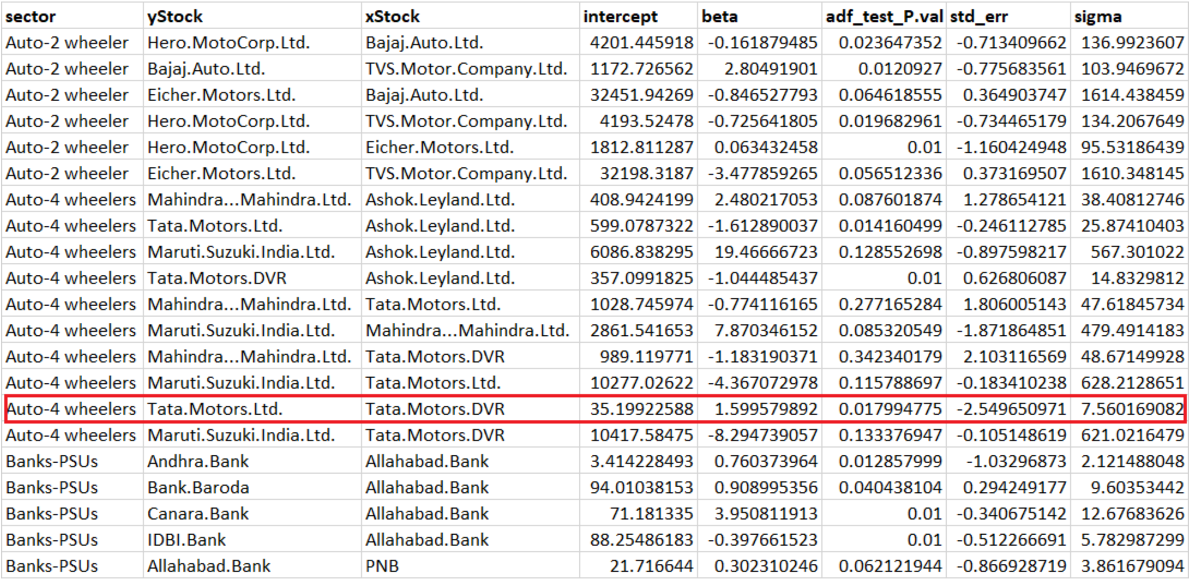
The sigma value in the pair data report is simply the standard error of the residual, which in the above case is 22.776.

So now if you read through the pair data sheet, you should be able to understand the details completely.

Alright, let us jump to the trade now

13.3 – Live example

I have been running the pair trading algo to look for opportunities, and I found one on 10th May 2018. Here is the snapshot of the pair data, you can download the same towards the end of this chapter. Do recollect, this pair trading algo was generated using the closing prices of 10th May.



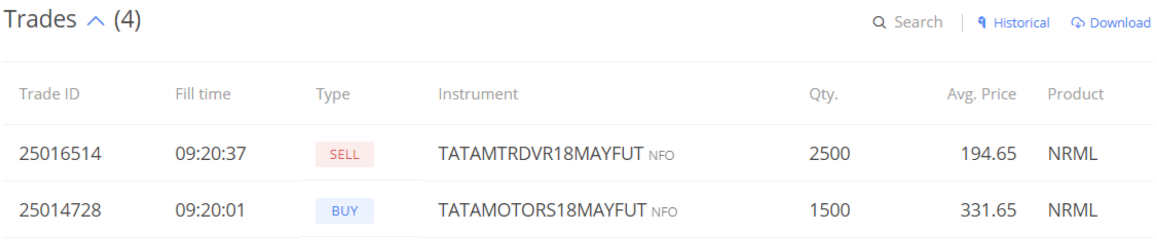
sector	yStock	xStock	intercept	beta	adf_test_P.val	std_err	sigma
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Bajaj.Auto.Ltd.	4201.445918	-0.161879485	0.023647352	-0.713409662	136.9923607
Auto-2 wheeler	Bajaj.Auto.Ltd.	TVS.Motor.Company.Ltd.	1172.726562	2.80491901	0.0120927	-0.775683561	103.9469672
Auto-2 wheeler	Eicher.Motors.Ltd.	Bajaj.Auto.Ltd.	32451.94269	-0.846527793	0.064618555	0.364903747	1614.438459
Auto-2 wheeler	Hero.MotoCorp.Ltd.	TVS.Motor.Company.Ltd.	4193.52478	-0.725641805	0.019682961	-0.734465179	134.2067649
Auto-2 wheeler	Hero.MotoCorp.Ltd.	Eicher.Motors.Ltd.	1812.811287	0.063432458	0.01	-1.160424948	95.53186439
Auto-2 wheeler	Eicher.Motors.Ltd.	TVS.Motor.Company.Ltd.	32198.3187	-3.477859265	0.056512336	0.373169507	1610.348145
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Ashok.Leyland.Ltd.	408.9424199	2.480217053	0.087601874	1.278654121	38.40812746
Auto-4 wheelers	Tata.Motors.Ltd.	Ashok.Leyland.Ltd.	599.0787322	-1.612890037	0.014160499	-0.246112785	25.87410403
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Ashok.Leyland.Ltd.	6086.838295	19.46666723	0.128552698	-0.897598217	567.301022
Auto-4 wheelers	Tata.Motors.DVR	Ashok.Leyland.Ltd.	357.0991825	-1.044485437	0.01	0.626806087	14.8329812
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.Ltd.	1028.745974	-0.774116165	0.277165284	1.806005143	47.61845734
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Mahindra...Mahindra.Ltd.	2861.541653	7.870346152	0.085320549	-1.871864851	479.4914183
Auto-4 wheelers	Mahindra...Mahindra.Ltd.	Tata.Motors.DVR	989.119771	-1.183190371	0.342340179	2.103116569	48.67149928
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.Ltd.	10277.02622	-4.367072978	0.115788697	-0.183410238	628.2128651
Auto-4 wheelers	Tata.Motors.Ltd.	Tata.Motors.DVR	35.19922588	1.599579892	0.017994775	-2.549650971	7.560169082
Auto-4 wheelers	Maruti.Suzuki.India.Ltd.	Tata.Motors.DVR	10417.58475	-8.294739057	0.133376947	-0.105148619	621.0216479
Banks-PSUs	Andhra.Bank	Allahabad.Bank	3.414228493	0.760373964	0.012857999	-1.03296873	2.121488048
Banks-PSUs	Bank.Baroda	Allahabad.Bank	94.01038153	0.908995356	0.040438104	0.294249177	9.60353442
Banks-PSUs	Canara.Bank	Allahabad.Bank	71.181335	3.950811913	0.01	-0.340675142	12.67683626
Banks-PSUs	IDBI.Bank	Allahabad.Bank	88.25486183	-0.397661523	0.01	-0.512266691	5.782987299
Banks-PSUs	Allahabad.Bank	PNB	21.716644	0.302310246	0.062121944	-0.866928719	3.861679094

Look at the data highlighted in red. This is Tata Motors Ltd as Y (dependent) and Tata Motors DVR as X (independent).

The ADF value reads, 0.0179 (less than the threshold of 0.05), and I think this is an excellent ADF value. Do recollect, ADF value of less than 0.05 indicates that the residual is stationary, which is exactly what we are looking for.

The std_err reads -2.54, which means the residuals is close has diverged (sufficiently enough) away from the mean and therefore one can look at setting up a long trade. Since this is a long trade, one is required to buy the dependent stock (Tata Motors) and short the independent stock (Tata Motors DVR). This trade was supposed to be taken on 11th May Morning (Friday), but for some reason, I was unable to place the trade. However, I did take the trade on 14th May (Monday) morning at a slightly bad rate, nevertheless, the intention was to showcase the trade and not really chase the P&L.

Here are the trade execution details –



Trade ID	Fill time	Type	Instrument	Qty.	Avg. Price	Product
25016514	09:20:37	SELL	TATAMTRDVR18MAYFUT NFO	2500	194.65	NRML
25014728	09:20:01	BUY	TATAMOTORS18MAYFUT NFO	1500	331.65	NRML

You may have two questions at this point. Let me list them for you –

Question – Did I actually execute the trade without checking for prices? As in I didn't even look at what price the stocks, I didn't look at support, resistance, RSI etc. Is it not required?

Answer – No, none of that is required. The only thing that matters is where the residual is trading, which is exactly what I looked for.

Question – On what basis did I choose to trade 1 lot each? Why can't I trade 2 lots of TM and 3 lots of TMD?

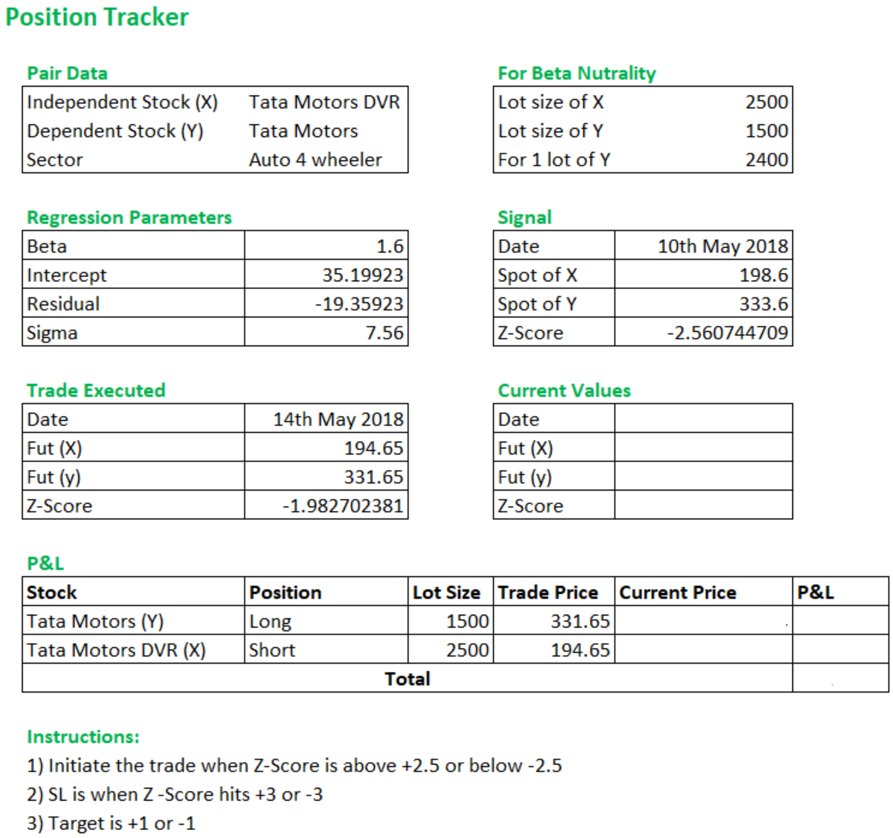
Answer – Well this depends on the beta of the stock. We will use the beta and identify the number of stocks of X &Y to ensure we are **beta neutral** in this position. The beta neutrality states that for every 1 stock of Y, we need to have beta*X stock of X. For example, in the Tata Motors (Y) and Tata Motors DVR (X) for example, the beta is 1.59. This means, for every 1 stock of Tata Motors (Y), I need to have 1.59 stocks of Tata Motors DVR (X).

Going by this proportion, the lot size of Tata Motors (Y) is 1500, so we need 1500*1.59 or 2385 shares of Tata Motors DVR (X). The lot size is 2400, quite close to 2385, hence I decided to go with 1 lot each. But I'm aware this trade is slightly more skewed towards the long side since I'm buying additional 115.

Also, please note, because of this constraint, we cannot really trade pairs if the beta is –ve, at least, not always.

Remember, I initiated this trade when the residual value was -2.54. The idea was to keep the position open and wait for the target (-1 on residual) or stop loss (-3 on residual) was hit. Until then, it was just a waiting game.

To track the position live, I've developed a basic excel tracker. Of course, if you are a programmer, you can do much better with these accessories, but given my limited abilities, I put up a basic position tracker in excel. Here is the snapshot, of course, you can download this sheet from the link posted below.



Position Tracker

Pair Data

Independent Stock (X)	Tata Motors DVR
Dependent Stock (Y)	Tata Motors
Sector	Auto 4 wheeler

For Beta Nutrality

Lot size of X	2500
Lot size of Y	1500
For 1 lot of Y	2400

Regression Parameters

Beta	1.6
Intercept	35.19923
Residual	-19.35923
Sigma	7.56

Signal

Date	10th May 2018
Spot of X	198.6
Spot of Y	333.6
Z-Score	-2.560744709

Trade Executed

Date	14th May 2018
Fut (X)	194.65
Fut (y)	331.65
Z-Score	-1.982702381

Current Values

Date	
Fut (X)	
Fut (y)	
Z-Score	

P&L

Stock	Position	Lot Size	Trade Price	Current Price	P&L
Tata Motors (Y)	Long	1500	331.65		
Tata Motors DVR (X)	Short	2500	194.65		
Total					

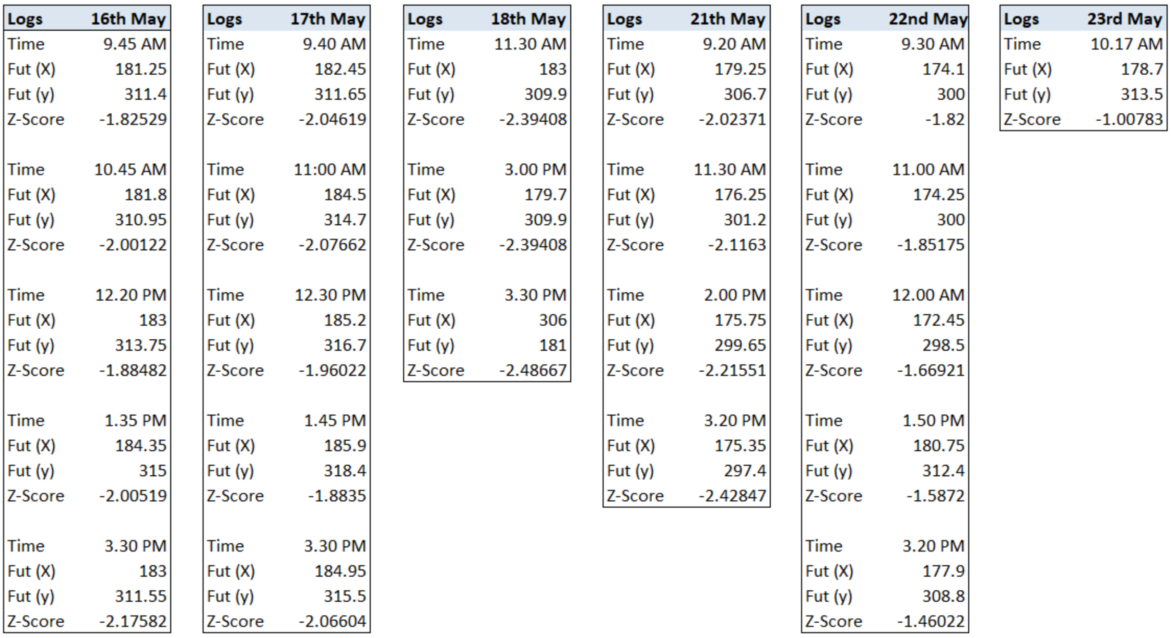
Instructions:

1) Initiate the trade when Z-Score is above +2.5 or below -2.5
2) SL is when Z -Score hits +3 or -3
3) Target is +1 or -1

The position tracker has all the basic information about the pair. I'm guessing this is a fairly easy sheet to understand. I've designed it in such a way that upon entering the current values of X & Y, the latest Z score is calculated and also the P&L. I'd encourage you

to play around this sheet, even better if you can build one yourself

Once the position is taken, all one has to do is track the z-score of the residual. This means you have to keep tracking the values and the respective z-scores. This is exactly what I did. In fact, for the sake of this chapter, my colleague, Faisal, logged all the values (except for the 14th and 15th). Here are the logs –



Logs	16th May
Time	9.45 AM
Fut (X)	181.25
Fut (y)	311.4
Z-Score	-1.82529
Time	10.45 AM
Fut (X)	181.8
Fut (y)	310.95
Z-Score	-2.00122
Time	12.20 PM
Fut (X)	183
Fut (y)	313.75
Z-Score	-1.88482
Time	1.35 PM
Fut (X)	184.35
Fut (y)	315
Z-Score	-2.00519
Time	3.30 PM
Fut (X)	183
Fut (y)	311.55
Z-Score	-2.17582

Logs	17th May
Time	9.40 AM
Fut (X)	182.45
Fut (y)	311.65
Z-Score	-2.04619
Time	11:00 AM
Fut (X)	184.5
Fut (y)	314.7
Z-Score	-2.07662
Time	12.30 PM
Fut (X)	185.2
Fut (y)	316.7
Z-Score	-1.96022
Time	1.45 PM
Fut (X)	185.9
Fut (y)	318.4
Z-Score	-1.8835
Time	3.30 PM
Fut (X)	184.95
Fut (y)	315.5
Z-Score	-2.06604

Logs	18th May
Time	11.30 AM
Fut (X)	183
Fut (y)	309.9
Z-Score	-2.39408
Time	3.00 PM
Fut (X)	179.7
Fut (y)	309.9
Z-Score	-2.39408
Time	3.30 PM
Fut (X)	306
Fut (y)	181
Z-Score	-2.48667

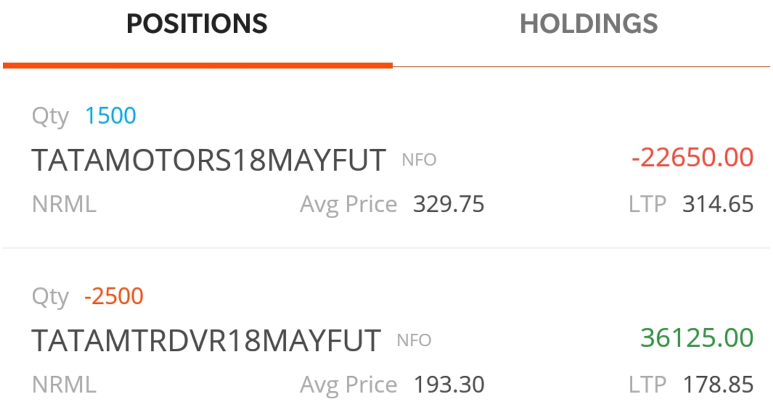
Logs	21th May
Time	9.20 AM
Fut (X)	179.25
Fut (y)	306.7
Z-Score	-2.02371
Time	11.30 AM
Fut (X)	176.25
Fut (y)	301.2
Z-Score	-2.1163
Time	2.00 PM
Fut (X)	175.75
Fut (y)	299.65
Z-Score	-2.21551
Time	3.20 PM
Fut (X)	175.35
Fut (y)	297.4
Z-Score	-2.42847

Logs	22nd May
Time	9.30 AM
Fut (X)	174.1
Fut (y)	300
Z-Score	-1.82
Time	11.00 AM
Fut (X)	174.25
Fut (y)	300
Z-Score	-1.85175
Time	12.00 AM
Fut (X)	172.45
Fut (y)	298.5
Z-Score	-1.66921
Time	1.50 PM
Fut (X)	180.75
Fut (y)	312.4
Z-Score	-1.5872
Time	3.20 PM
Fut (X)	177.9
Fut (y)	308.8
Z-Score	-1.46022

Logs	23rd May
Time	10.17 AM
Fut (X)	178.7
Fut (y)	313.5
Z-Score	-1.00783

As you can see, the current values were tracked and the latest z-score was calculated several times a day. The position was open for nearly 7 trading session and this is quite common with pair trading. I've experienced positions where they were open for nearly 22 - 25 trading sessions. But here is the thing – as long as your math is right, you just have to wait for the target or SL to trigger.

Finally, on 23rd May morning, the z-score dropped to the target level and there was a window of opportunity to close this trade. Here is the snapshot –

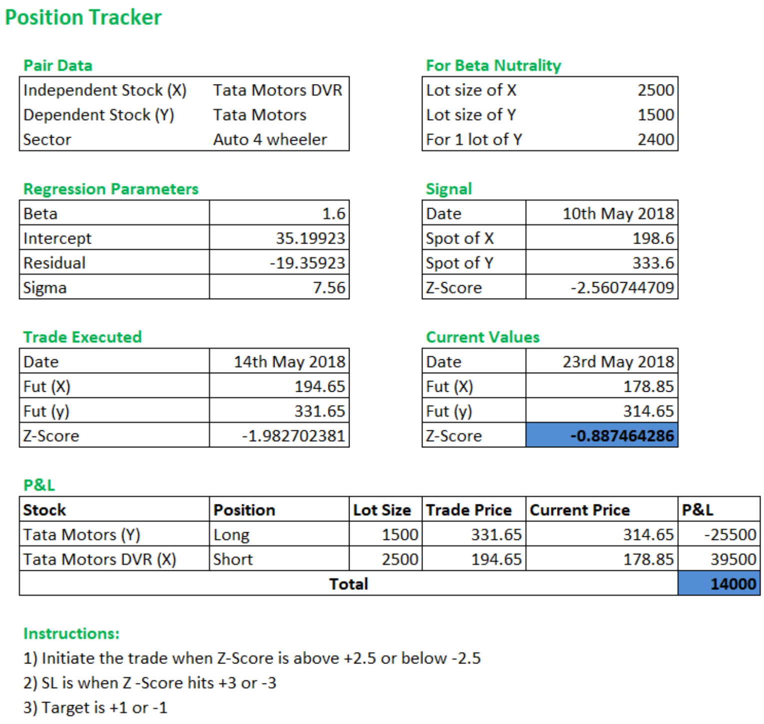


Qty 1500
TATAMOTORS18MAYFUT NFO -22650.00
NRML Avg Price 329.75 LTP 314.65

Qty -2500
TATAMTRDVR18MAYFUT NFO 36125.00
NRML Avg Price 193.30 LTP 178.85

Notice, the gains in Tata Motors DVR is much larger than the loss in Tata Motors. In fact, when we take the trade, we will never know which of the two positions will make us the money. The idea, however, is that one of them will move in our favour and the other won't (or may). It's however, just not possible to identify which one will be the breadwinner.

The position tracker for the final day (23rd May) looked like this –



Position Tracker

Pair Data

Independent Stock (X)	Tata Motors DVR
Dependent Stock (Y)	Tata Motors
Sector	Auto 4 wheeler

For Beta Nutrality

Lot size of X	2500
Lot size of Y	1500
For 1 lot of Y	2400

Regression Parameters

Beta	1.6
Intercept	35.19923
Residual	-19.35923
Sigma	7.56

Signal

Date	10th May 2018
Spot of X	198.6
Spot of Y	333.6
Z-Score	-2.560744709

Trade Executed

Date	14th May 2018
Fut (X)	194.65
Fut (y)	331.65
Z-Score	-1.982702381

Current Values

Date	23rd May 2018
Fut (X)	178.85
Fut (y)	314.65
Z-Score	**-0.887464286**

P&L

Stock	Position	Lot Size	Trade Price	Current Price	P&L
Tata Motors (Y)	Long	1500	331.65	314.65	-25500
Tata Motors DVR (X)	Short	2500	194.65	178.85	39500
Total					**14000**

Instructions:

1) Initiate the trade when Z-Score is above +2.5 or below -2.5

2) SL is when Z -Score hits +3 or -3

3) Target is +1 or -1

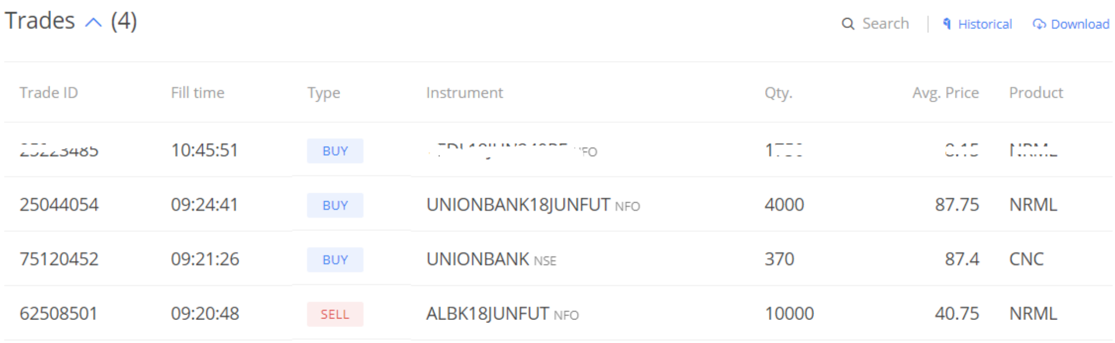
The P&L was roughly Rs.14,000/-, not bad I'd say for a relatively low-risk trade.

13.4 – Final words on Pair Trading

Alright guys, over the last 13 chapter, we have discussed everything I know about pair trading. I personally thing this is a very exciting way of trading rather than blind speculative trading. Although less risky, pair trade has its own share of risk and you need to be aware of the risk. One of the common ways to lose money is when the pair can continue to diverge after you initiate the position, leaving you with a deep loss. Further, the margin requirements are slightly higher since there are two contracts you are dealing with. This also means you need to have some buffer money in your account to accommodate daily M2M.

There could be situations where you will need to take a position in the spot market as well. For example, on 23rd May, there was a signal to go short on Allahabad Bank (Y) and long on Union Bank (X). The z-score was 2.64 and the beta for this pair is 0.437.

Going by beta neutrality, for every 1 share of Allahabad Bank (Y), I need 0.437 shares of Union Bank (X). The Lot size of Allahabad Bank is 10,000, this implies I need to buy 4378 shares of Union Bank. However, the lot size of Union Bank is 4000, hence I had to buy 370 shares in the spot market.



Trade ID	Fill time	Type	Instrument	Qty.	Avg. Price	Product
25223485	10:45:51	BUY	[illegible] NFO	[illegible]	[illegible]	NRML
25044054	09:24:41	BUY	UNIONBANK18JUNFUT NFO	4000	87.75	NRML
75120452	09:21:26	BUY	UNIONBANK NSE	370	87.4	CNC
62508501	09:20:48	SELL	ALBK18JUNFUT NFO	10000	40.75	NRML

Well, I hope I trade is successful

I know most of you would want the pair data sheet made available. We are working on making this sheet available to you on a daily basis so that you can track the pairs. Meanwhile, I would suggest you try and build this algo yourself. If you have concerns, please post it below and I will be happy to assist.

If you don't know how to program then you have no option but to find someone who knows programming and convince him or her that there is money to be made, this is

exactly what I did

Lastly, I would like to leave you with a thought –

1. We run a linear regression of Stock A with Stock B to figure out if the two stocks are cointegrated with their residuals being stationary
2. What if Stock A with Stock B is not stationary, but instead Stock A is stationary with stock B & C as a combined entity?

Beyond Pair, trading lies something called as multivariate regression. By no stretch of the imagination is this easy to understand, but let me tell you if you can graduate to this arena, the game is different.

Download the Position Tracker and Pair Datasheet below:

DOWNLOAD POSITION TRACKER

DOWNLOAD PAIR DATASHEET

Key takeaways from this chapter

1. The trigger to trade a pair comes from the residual’s current value
2. Check for beta neutrality of the pair to identify the number of stock required in X and Y
3. If the beta of the pair is negative, then it may not be possible to set up the trade
4. Once the trade is initiated, check the z-score movement to trade its current position
5. The price of the futures does not really matter, the emphasis is only on the z-score

CHAPTER 14

Live Example – 2

14.1 – Position Sizing

I know, the discussion on pair trading was to end with the previous chapter, but I thought I had to discuss a special case before we finally wrap up. I'll also try and keep this chapter

really short

So here you go.

I ran through the pair trading algo y'day evening (28th May) and found a very interesting trade. Here are the regression parameters –

- Stock X = ICICI Bank
- Stock Y = HDFC Bank
- ADF = 0.048
- Beta = 0.79
- Intercept = 1626
- Std_err = 2.67

What do you think of it? Perfect isn't it? Its ICICI and HDFC, two of the largest private sector banks, both have similar business landscape, both have a similar revenue stream, both regulated by RBI. Perhaps the perfect candidate for a pair trade, right?

The ADF value is 0.048, which means there is only 4.8% chance that the residual is non-stationary or about 95.2% chance of the residuals being stationary, which is fantastic.

The std_err is +2.67, which is a perfect residual value to initiate a short pair trade. The trade here is short HDFC and go long on ICIC.

So, how do we position size this? Here are the price and lot size details –

- HDFC Fut Price = 2024.8
- HDFC Lot size = 500
- ICICI Fut price = 298.8
- ICICI Lot size = 2750

Remember we discussed position size in the previous chapter. We look at the beta and estimate the number of shares required for this trade.

The beta is 0.79, this means, every 1 share of Y needs to be offset with 0.79 shares of X. The lot size of HDFC (Y) is 500, this means to offset the beta, we need 395 shares of ICICI (X).

Do you see the problem here? The lot sizes simply do not match.

We cannot simply trade 1 lot each here like we did in the TATA Motors and Tata Motors DVR example, discussed in the previous chapter. If we do, then this won't be a beta neutral trade.

Hence to position size this, we need to work around with the lot sizes –

The lot size of ICIC is 2750, beta is 0.79, lot size of HDFC is 500. Given this, that the lot size is higher than HDFC, what should be the minimum number of HDFC shares which will beta neutral 2750 shares of ICICI.

To figure this out, we simply divide –

2750/0.79

= 3481.01

Since the lot size of HDFC is 500, we can round this off to 3500. Considering the lot size of HDFC is 500, this will be 7 lots of HDFC against 1 lot of ICICI.

14.2 – Intercept

Alright, now that we know the position size as well, here is the big question – will you take this trade?

Everything seems perfect, right? ADF has a desirable value, residual is at 2.67 SD, the two stocks are highly correlated, the business is similar. So what can go wrong?

Yes, I agree, everything looks good, but on a closer look, the intercept reveals a slightly different story.

To understand this, we need to quickly revisit the regression equation –

y = Beta * x + Intercept + Residual

If you think about this equation, we are trying to explain the stock price of Y in terms of the stock price of X multiplied by its beta. The intercept is essentially that portion of the y's stock price which the model cannot explain, and the residual is the difference between predicted y and actual y.

Going by this, a large intercept implies that a large portion of Y's stock price cannot be explained by the regression model.

In this case, the intercept is 1626. The stock price of HDFC is 2024 per share, this means, 1626 out of 2024 cannot be explained by the regression equation. This means, the regression equation cannot explain nearly 80% (1626/2024) of Y's stock price or in other words the equation can explain only 20% of the equation, which according to me is quite tricky.

This further implies, that if we are trading this pair, then we are essentially trading a very small probability here. I'd rather avoid this and look for another opportunity than trade this. Of course, I know traders who would love to jump in and take this trade, but for

someone like me, I'd look at risk first and then the reward

Good luck!

DOWNLOAD PAIR DATASHEET

CHAPTER 15

Calendar Spreads

15.1 – The classic approach

I had briefly introduced the concept of calendar spreads in **Chapter 10** of the Futures Trading module. Traditionally calendar spreads are dealt with a price based approach. Here is a quick recap on how this is done –

1. Calculate the fair value of current month contract
2. Calculate the fair value of the mid-month contract
3. Look for relative mispricing between the two contracts

Based on the mispricing, you either buy the current month contract and sell the mid-month contract or sell the current month contract and buy the mid-month contract. Here is an example of a Calendar Spread –

1. Buy TCS Futures expiring 28th June 2018 @ 1846
2. Sell TCS Futures expiring 28th July 2018 @ 1851

Here you buy and sell the futures of the same stock, but of contracts belonging to different expiries like showcased above. The difference between prices of the two contracts is what is expected to made here. The risk is extremely low in calendar spreads so therefore the money you make on calendar spreads is also small. If you are trader like me, who is averse to risk, then this is something you may like.

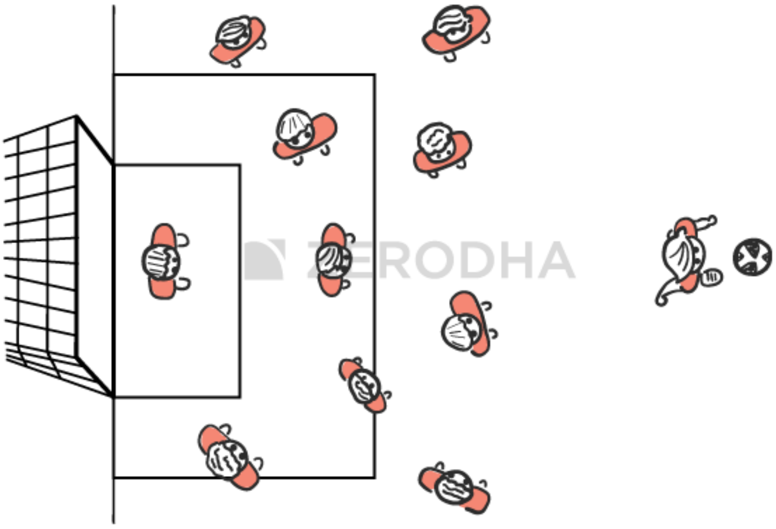
This approach to performing a calendar spread is a decent one.

By the way, if you are not familiar with what I'm discussing, then I'd suggest you read Chapter 10 in the Futures Trading module to get a quick perspective on the classic calendar spreads approach. I think it forms a crucial foundation on top of which you can build other variant/styles of calendar spreads.

So let's get started straight away.

15.2 – Calendar spread logic

If you have read the chapters on pair trading, then understanding the calendar spread logic is quite straightforward. This simplified approach assumes that the current price of futures is a reflection of everything known in the market. The known set of information can extend from news on the stock, corporate action, discount/premium, fair value, and literally everything out there which is relevant to the stock.



Now, if the above assumption is valid, then probably we can use the price itself as a trigger to identify opportunities to set up a calendar spread trade. This kind of simplifies the whole approach. Calendar spreads are a low-risk strategy so therefore do not expect big bucks from this strategy. However, since you simultaneously buy-sell the same asset, you take out the directional risk involved in the trade, hence it does make sense to top up the leverage. Also, unlike pair trade, the calendar spread trades can be ultra-short term in nature, with most of the trades closing within the same day. Before I take up an example to explain this, I'll quickly give you an overview of this is done.

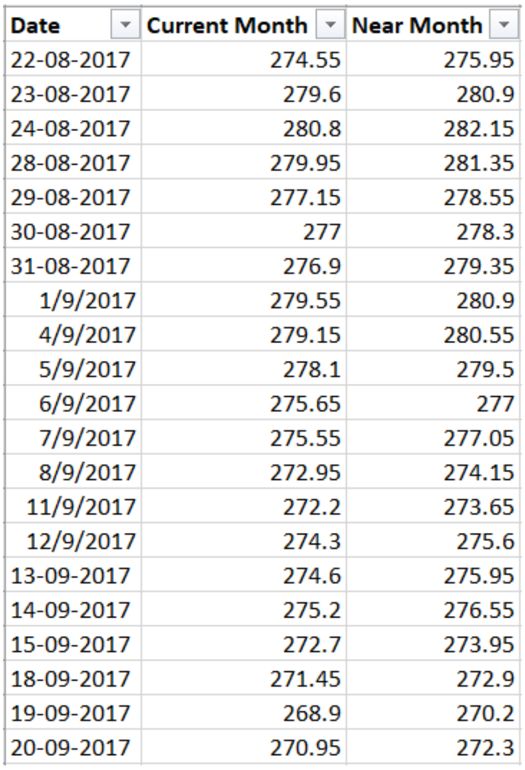
Start with downloading the continuous futures closing prices of the stock for both near month and next month contracts.

Calculate the daily historic difference between the two contracts and generate a time series. Calculate the mean and standard deviation of the time series. Using the mean and standard deviation data we can estimate the range for the difference. A trading signal is triggered when the difference between the two contracts move to mean plus or minus 1 standard deviation and the trade is closed when the difference collapses to mean.

You get the point, don't you

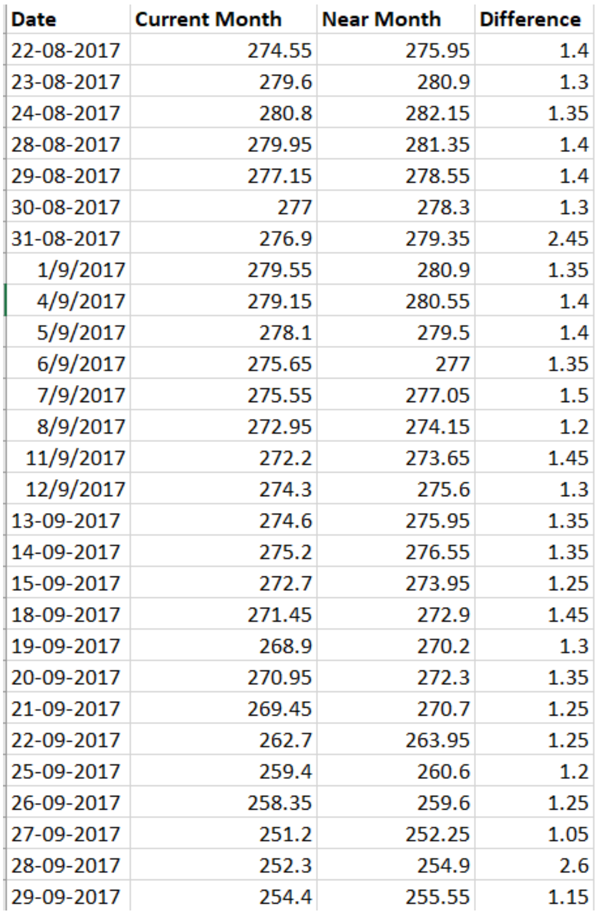
15.3 – Calendar spread example

I've taken the example of SBIN to illustrate calendar spreads. I have download the continuous futures data from Zerodha Pi (Zerodha's desktop trading application) for last 200 trading days. I have got the closing prices on excel sheet, and this is how it looks –



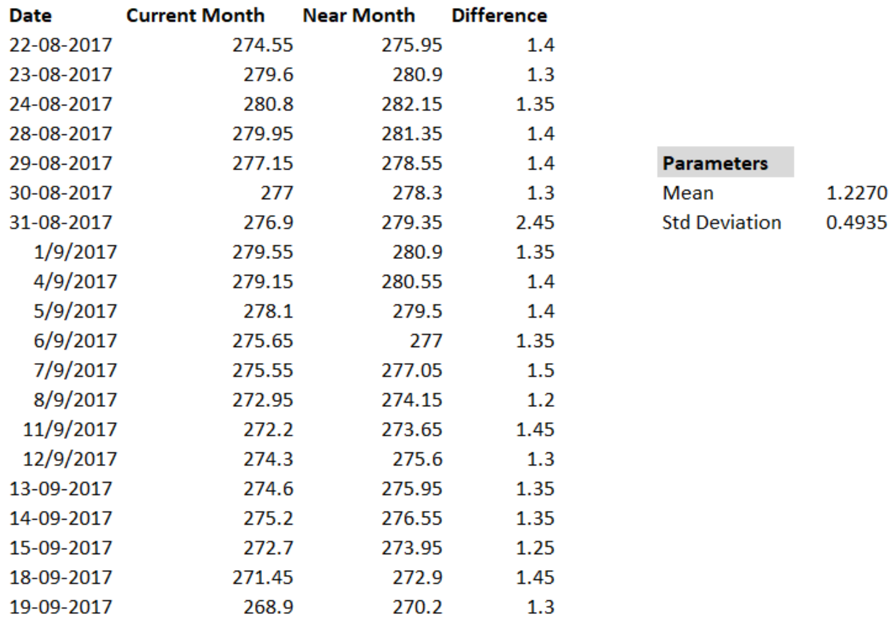
Date	Current Month	Near Month
22-08-2017	274.55	275.95
23-08-2017	279.6	280.9
24-08-2017	280.8	282.15
28-08-2017	279.95	281.35
29-08-2017	277.15	278.55
30-08-2017	277	278.3
31-08-2017	276.9	279.35
1/9/2017	279.55	280.9
4/9/2017	279.15	280.55
5/9/2017	278.1	279.5
6/9/2017	275.65	277
7/9/2017	275.55	277.05
8/9/2017	272.95	274.15
11/9/2017	272.2	273.65
12/9/2017	274.3	275.6
13-09-2017	274.6	275.95
14-09-2017	275.2	276.55
15-09-2017	272.7	273.95
18-09-2017	271.45	272.9
19-09-2017	268.9	270.2
20-09-2017	270.95	272.3

The next step is to calculate the difference between the two contracts. It is advisable to subtract the price of near month contract from the current month contract. This is because, all else equal, the futures price of Near month contract is always higher than the previous month contract owing to the 'cost of carry'. Chapter 10 of futures module explains this in more detail. The difference is calculated and the time series data is generated, as shown below –



Date	Current Month	Near Month	Difference
22-08-2017	274.55	275.95	1.4
23-08-2017	279.6	280.9	1.3
24-08-2017	280.8	282.15	1.35
28-08-2017	279.95	281.35	1.4
29-08-2017	277.15	278.55	1.4
30-08-2017	277	278.3	1.3
31-08-2017	276.9	279.35	2.45
1/9/2017	279.55	280.9	1.35
4/9/2017	279.15	280.55	1.4
5/9/2017	278.1	279.5	1.4
6/9/2017	275.65	277	1.35
7/9/2017	275.55	277.05	1.5
8/9/2017	272.95	274.15	1.2
11/9/2017	272.2	273.65	1.45
12/9/2017	274.3	275.6	1.3
13-09-2017	274.6	275.95	1.35
14-09-2017	275.2	276.55	1.35
15-09-2017	272.7	273.95	1.25
18-09-2017	271.45	272.9	1.45
19-09-2017	268.9	270.2	1.3
20-09-2017	270.95	272.3	1.35
21-09-2017	269.45	270.7	1.25
22-09-2017	262.7	263.95	1.25
25-09-2017	259.4	260.6	1.2
26-09-2017	258.35	259.6	1.25
27-09-2017	251.2	252.25	1.05
28-09-2017	252.3	254.9	2.6
29-09-2017	254.4	255.55	1.15

I will now calculate the mean and standard deviation on this time series. The mean will give me an estimate on how much of the difference is acceptable on a ‘day to day’ basis and at the same time, the standard deviation will give me a sense of variation in this difference. Here is the snapshot.



Date	Current Month	Near Month	Difference
22-08-2017	274.55	275.95	1.4
23-08-2017	279.6	280.9	1.3
24-08-2017	280.8	282.15	1.35
28-08-2017	279.95	281.35	1.4
29-08-2017	277.15	278.55	1.4
30-08-2017	277	278.3	1.3
31-08-2017	276.9	279.35	2.45
1/9/2017	279.55	280.9	1.35
4/9/2017	279.15	280.55	1.4
5/9/2017	278.1	279.5	1.4
6/9/2017	275.65	277	1.35
7/9/2017	275.55	277.05	1.5
8/9/2017	272.95	274.15	1.2
11/9/2017	272.2	273.65	1.45
12/9/2017	274.3	275.6	1.3
13-09-2017	274.6	275.95	1.35
14-09-2017	275.2	276.55	1.35
15-09-2017	272.7	273.95	1.25
18-09-2017	271.45	272.9	1.45
19-09-2017	268.9	270.2	1.3

Parameters	
Mean	1.2270
Std Deviation	0.4935

You can calculate the mean and standard deviation on excel using the '=Average ()' and '=stdev()' functions respectively.

The mean of 1.227 tells me that, all else equal, the difference between the two contracts should be 1.227 or in that vicinity. This essentially means, there is no trade opportunity if the spread (or the difference) between the two contracts hovers around this value.

We now use the standard deviation value and the mean value to calculate the range of the spread –

- Upper range = 1.227 + 0.4935 = 1.7205
- Lower Range = 1.227 – 0.4935 = 0.7335

I had mentioned that the spread can hover around 1.227, but I had not quantified 'vicinity', which is quite important. The range calculation does just that, it helps us quantify the range within which (vicinity) the spread can vary on a daily basis. Any value of the spread outside this range gives us an opportunity to set up a calendar spread.

If the spread has increased beyond the upper range of 1.7205, it means either the near month contract has increased in value or the current month contract has reduced in value.

The rule of thumb in any arbitrage is to always buy the asset in the cheaper market and sell the same asset in the expensive market, hence the trade here would be to **buy the current month contract and sell the near month contract**.

Likewise, if the spread has fallen below the lower range value i.e. 0.7335, this means the current month has become expensive and near month has become cheaper. Hence, the trade here is to **sell the current month and buy the near month contract**.

With this logic in perspective, let's evaluate the if SBIN has given us any opportunities over the last 200 trading days.

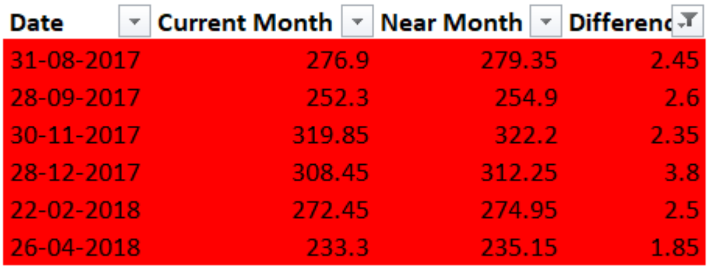
15.4 – Spotting opportunities

Keeping the above pointers in perspective, we can conclude the following –

1. Sell the spread when the spread increases beyond 1.7205. Sell spread means, sell the near month contract and buy the current month contract
2. Buy the spread when the spread shrinks below 0.7335. Buy spread means, buy the near month contract and sell the current month contract.

If you find it hard to figure out which contract to buy and which one to sell when a signal originates, then simply think in terms of the near month contract. Sell spread means sell the near month (therefore buy current month) and buy spread means buy the near-month (therefore sell the current month contract).

In the excel sheet, I now look for the historical opportunities. I will identify the sell spread opportunities first. To do this, I simply have to apply a filter, to filter out all values above 1.7205. I've done the same, here are the results –



Date	Current Month	Near Month	Differenc
31-08-2017	276.9	279.35	2.45
28-09-2017	252.3	254.9	2.6
30-11-2017	319.85	322.2	2.35
28-12-2017	308.45	312.25	3.8
22-02-2018	272.45	274.95	2.5
26-04-2018	233.3	235.15	1.85

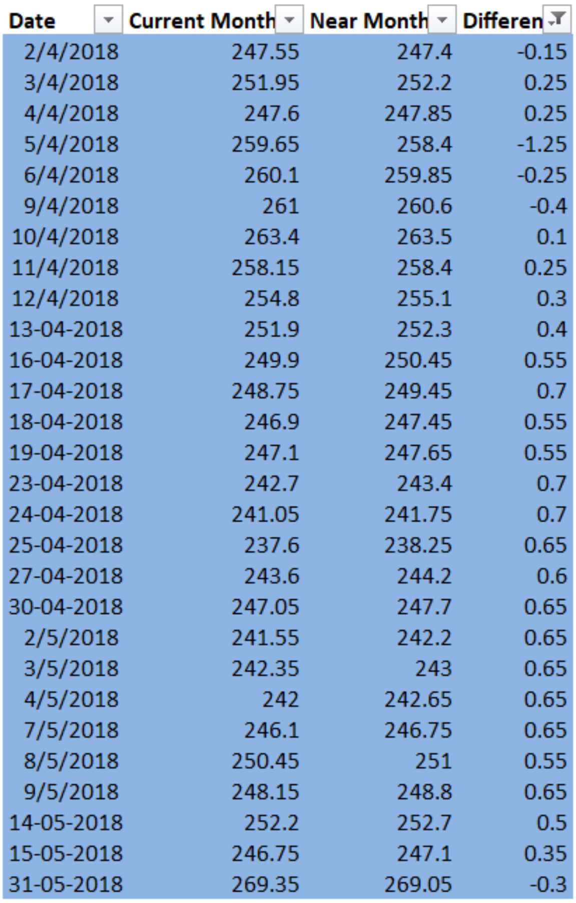
As you can see, on 6 occasions, the spread increases beyond 1.7205 or the first standard deviation levels. On all these occasions, there was a trigger to sell, implying the spread would fall back to mean.

In fact, here is how the spread behaved –

Signal Date	Sell spread value	Trade closing date	Buy spread value	P&L
31-08-2017	2.45	1-09-2017	1.35	1.1
28-092017	2.6	29-09-2017	1.15	1.45
30-11-2017	2.35	01-12-2017	1.55	0.8
28-12-2012	3.8	29-12-2017	1.45	2.35
22-02-2018	2.5	23-03-2018	1.3	1.2
26-04-2018	1.85	27-04-2018	0.6	1.25

As you can notice, signals originate around month ends, probably due to expiry dynamics. Also, every trade has resulted in a profit (although small) and closed the very next day.

Let us see how the buy spread trades have performed. I have filtered for all values below 0.7335, and here are the results –



Date	Current Month	Near Month	Differen
2/4/2018	247.55	247.4	-0.15
3/4/2018	251.95	252.2	0.25
4/4/2018	247.6	247.85	0.25
5/4/2018	259.65	258.4	-1.25
6/4/2018	260.1	259.85	-0.25
9/4/2018	261	260.6	-0.4
10/4/2018	263.4	263.5	0.1
11/4/2018	258.15	258.4	0.25
12/4/2018	254.8	255.1	0.3
13-04-2018	251.9	252.3	0.4
16-04-2018	249.9	250.45	0.55
17-04-2018	248.75	249.45	0.7
18-04-2018	246.9	247.45	0.55
19-04-2018	247.1	247.65	0.55
23-04-2018	242.7	243.4	0.7
24-04-2018	241.05	241.75	0.7
25-04-2018	237.6	238.25	0.65
27-04-2018	243.6	244.2	0.6
30-04-2018	247.05	247.7	0.65
2/5/2018	241.55	242.2	0.65
3/5/2018	242.35	243	0.65
4/5/2018	242	242.65	0.65
7/5/2018	246.1	246.75	0.65
8/5/2018	250.45	251	0.55
9/5/2018	248.15	248.8	0.65
14-05-2018	252.2	252.7	0.5
15-05-2018	246.75	247.1	0.35
31-05-2018	269.35	269.05	-0.3

There are close to 28 trade here and not all of them are successful. Of course, the losses are as small as the profits, if not smaller. I'll let you do the exact calculation; like the way I've shown for the short trades.

I hope this example gives you a general sense of how to carry out calendar spread. I'm sure you'd agree that this is far simpler and intuitive compared to the classic approach to calendar spreads.

I have summarized my thoughts on Calendar spreads here and this will also double up as the key takeaways for this chapter –

1. The expected profits and losses are small in calendar spreads
2. Directional risk is eliminated; hence you go can go full throttle on leverage
3. All the short trades in SBIN were successful but longs were not – this implies that I would only look for short opportunities in SBI. In other words, you need to backtest

the P&L profile of each futures contract and figure out which contract you can go long on and which contract you can go short on

4. Since the P&L is small, ensure your trading costs are minimum, a discount broker like Zerodha is most suited for such trades J
5. Trades usually close within a day or two
6. Trades usually originate around expiry due to expiry dynamics

Think about this, if you can backtest this across the entire universe of equity and commodities futures contract, you will essentially have at least a signal or 2 every day!

I'd love to hear your thoughts, so please do post your queries.

DOWNLOAD THE EXCEL SHEET

PS: I won't be posting any new chapters for a while, but that does not mean I'm not working on new content, it is just that the delivery format will be different and way more exciting!

Stay tuned

CHAPTER 16

Momentum Portfolios

16.1 – Defining Momentum

If you have spent some time in the market, then I'm quite certain that you've been bombarded with market jargons of all sorts. Most of us get used to these jargons and in fact, start using these jargons without actually understanding what they really mean. I'm guilty of using few jargons without understanding the true meaning of it and I get a feeling that some of you reading this may have experienced the same.

One such jargon is – momentum. I'm sure we have used momentum is our daily conversations related to the markets, but what exactly is momentum and how is it measured?

When asked, traders loosely define momentum as the speed at which the markets move. This is correct to some extent, but that’s not all and we should certainly not limit our understanding to just that.

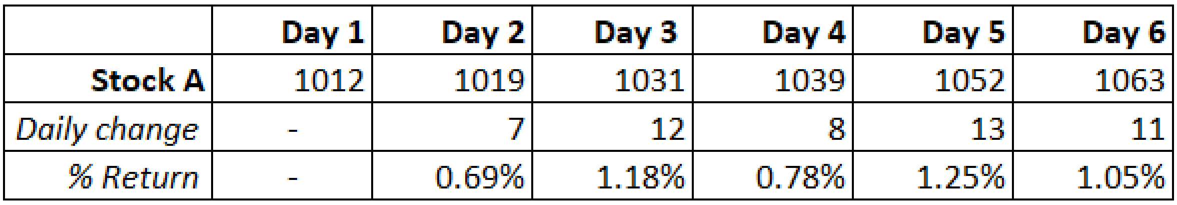
‘Momentum’ is a physics term, it refers to the quantity of motion that an object has. If you look at this definition in the context of stocks markets, then everything remains the same, except that you will have to replace ‘object’ by stocks or the index.

Simply put, momentum is the rate of change of returns of the stock or the index. If the rate of change of returns is high, then the momentum is considered high and if the rate of change of returns is low, the momentum is considered low.

This leads us to the next obvious question i.e. is what is the rate of change of returns?

The rate of change of return, as it states the return generated (or eroded) between two reference time period. For the sake of this discussion, let’s stick to the rate of change of return on an end of day basis. So in this context, the rate of change of returns simply means the speed at which the daily return of the stock varies.

To understand this better, consider this example –

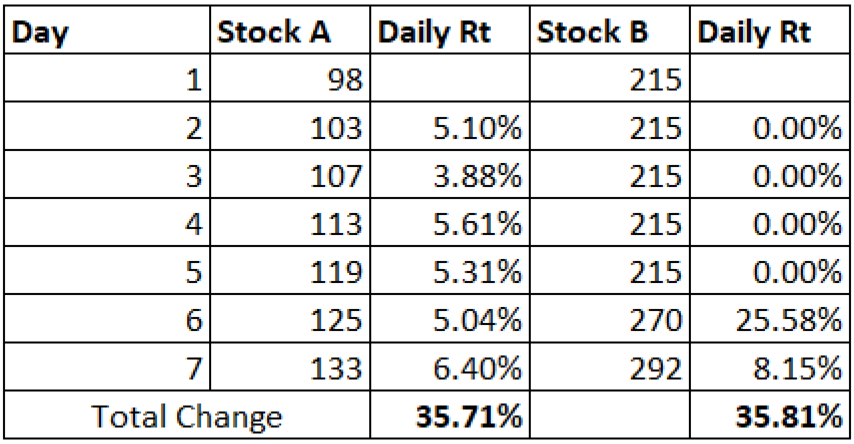


	Day 1	**Day 2**	**Day 3**	**Day 4**	**Day 5**	**Day 6**
Stock A	1012	1019	1031	1039	1052	1063
Daily change	-	7	12	8	13	11
% Return	-	0.69%	1.18%	0.78%	1.25%	1.05%

The table above shows the daily stock closing price of an arbitrary stock for 6 days. Two things to note here –

- The prices are moving up on day to day basis
- The percentage change is 0.5% or higher on a daily basis

Consider another example –



Day	Stock A	Daily Rt	Stock B	Daily Rt
1	98		215	
2	103	5.10%	215	0.00%
3	107	3.88%	215	0.00%
4	113	5.61%	215	0.00%
5	119	5.31%	215	0.00%
6	125	5.04%	270	25.58%
7	133	6.40%	292	8.15%
Total Change		**35.71%**		**35.81%**

Two things need to note –

- The prices are moving up on day to day basis
- The percentage change is 1.5% or higher on a daily basis

Given the behaviour of these two stocks, I have two questions for you –

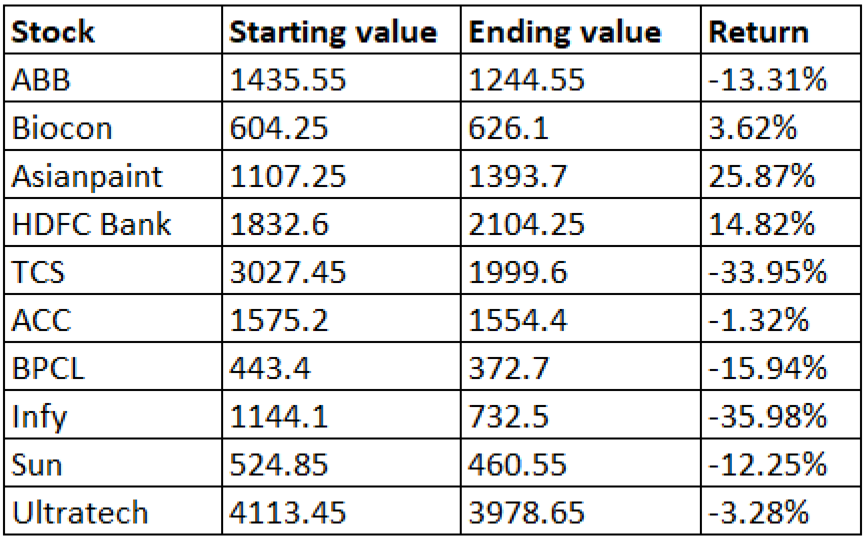
- Which stock has a higher rate of change in daily returns?
- Which sock has a higher momentum?

To answer these above questions, you can look at either the absolute change in Rupee value or the percentage change from a close to close perspective.

If you look at the absolute Rupee change, then obviously the change in Stock A is higher than Stock B. However, this is not the right way to look at the change in daily return. For instance, in absolute Rupee terms, stock in the range of say 2000 or 3000 will always have a higher change compared to Stock A.

Hence, evaluating absolute Rupee change will not suffice and therefore we need to look at the percentage change. In terms of percentage change, clearly Stock B's daily change is higher and therefore we can conclude that Stock B has a higher momentum.

Here is another situation, consider this –



Stock	Starting value	Ending value	Return
ABB	1435.55	1244.55	-13.31%
Biocon	604.25	626.1	3.62%
Asianpaint	1107.25	1393.7	25.87%
HDFC Bank	1832.6	2104.25	14.82%
TCS	3027.45	1999.6	-33.95%
ACC	1575.2	1554.4	-1.32%
BPCL	443.4	372.7	-15.94%
Infy	1144.1	732.5	-35.98%
Sun	524.85	460.55	-12.25%
Ultratech	4113.45	3978.65	-3.28%

Stock A, has trended up consistently on a day to day basis, while stock B has been quite a dud all along except for the last two days. On an overall basis if you check the percentage change over the 7-day period then both have delivered similar results. Given this, which of these two stock is considered to have good momentum?

Well, clearly Stock A is consistent in terms of daily returns, exhibits a good uptrend, and therefore can be considered to have continuity in showcasing momentum.

Now, what if I decide to measure momentum slightly differently? Instead of daily returns, what if we were to look at the return on a 7 days' basis? If we were to do that, then both Stock A and B would qualify as momentum stocks.

The point that I'm trying to make here is that traders generally tend to look at momentum in terms of daily returns, which is perfectly valid, but this is not necessarily the only way to look at momentum. In fact, the momentum strategy we will discuss later in this chapter looks at momentum on a larger time frame and not no daily basis. More on this later.

I hope by now, you do have a sense of what exactly momentum means and understood the fact that momentum can be measured not just in terms of daily returns but also in terms of larger time frames. In fact, high-frequency traders measure momentum on a minute to minute or hourly basis.

16.2 – Momentum Strategy

Amongst the many trading strategies that the traders use, one of the most popular strategies is the momentum strategy. Traders measure momentum in many different ways to identify opportunity pockets. The core idea across all these strategies remains the same i.e. to identify momentum and ride the wave.

Momentum strategies can be developed on a single stock basis wherein the idea is to measure momentum across all the stocks in the tracking universe and trade the ones which showcase the highest momentum. Do note, momentum can be either way – long or short, so a trader following single stock momentum strategy will get both long and short trading opportunities.

Traders also develop momentum strategies on a sector-specific basis and set up sector-specific trades. The idea here is to identify sector which exhibits strong momentum, this can be done by checking momentum in sector-specific indices. Once the sector is identified, further look for the stocks within the sector which display maximum strength in terms of momentum.

Momentum can also be applied on a portfolio basis. This involves the concept of portfolio creation with say 'n' number of stock, with each stock in the portfolio showcasing momentum. In my opinion, this is a great strategy as it is not just plain vanilla momentum strategy but also offers safety in terms of diversification.

We will discuss one such strategy wherein the idea is to create a basket of stock aka a portfolio consisting of 10 momentum stocks. Once created, the portfolio is held until the momentum lasts and then re-balanced.

16.3 – Momentum Portfolio

Before we discuss this strategy, I want you to note a few things –

- The agenda here is to highlight how a momentum portfolio can be set up. However, this is not the only way to build a momentum portfolio

- You will need programming skills to implement this strategy or to build any other momentum strategy. If you are not a coder like me, then do find a friend who can help
- Like any other strategy, this too has to be backtested

Given the above, here is a systematic guide to building a ‘Momentum Portfolio’.

Step 1 – Define your stock universe

As you may know, there are close to 4000 listed stocks on BSE and about 1800 on NSE. This includes highly valuable companies like TCS and absolute thuds such as pretty much all the Z category stocks on BSE. Companies such as these form the two extreme ends of the spectrum. The question is, do have to track all these stocks to build a momentum portfolio?

Not really, doing so would be a waste of time.

One has to filter out the stocks and create something called as the ‘tracking universe’. The tracking universe will consist of a large basket of stocks within which we will pick stocks to constitute the momentum portfolio. This means the momentum portfolio will always be a subset of the tracking universe.

Think of the tracking universe as a collection of your favourite shopping malls. Maybe out of the 100s of malls in your city, you may end up going to 2-3 shopping malls repeatedly. Clothes bought from these 2-3 malls make up for your entire wardrobe (read portfolio). Hence, these 2-3 malls end up forming your tracking universe out of the 100s available in your city.

The tracking universe can be quite straightforward – it can be the Nifty 50 stocks or the BSE 500 stocks. Therefore, the momentum portfolio will always be a subset of either the Nifty 50 or BSE 500 stocks. Keeping the BSE 500 stocks as your tracking universe is a good way to start, however, if you feel a little adventurous, you can custom create your tracking universe.

Custom creation can be on any parameter – for example, out of the entire 1800 stocks on NSE, I could use a filter to weed out stocks, which has a market cap of at least 1000Crs. This filter alone will shrink the list to a much smaller, manageable set. Further, I may add other criteria such as the price of the stock should be less than 2000. So on and so forth.

I am just randomly sharing few filter ideas, but you get the point. Using the custom creation techniques helps you filter out and build a tracking universe that exactly matches your requirement.

Lastly, from my personal experience, I would suggest you have at least 150-200 stocks in your tracking universe if you wish to build a momentum portfolio of 12-15 stock.

Step 2 – Set up the data

Assuming your tracking universe is set up, you are now good to proceed to the 2nd step. In this step, you need to ensure you get the closing prices of all the stocks in your tracking universe. Ensure the data set that you have is clean and adjusted for corporate actions like the bonus issue, splits, special dividends, and other corporate actions. Clean data is the key building block to any trading strategy. There are plenty of data sources from where you can download the data free, including the NSE/BSE websites.

The question is – what is the lookback period? How many historical data points are required? To run this strategy, you only need 1-year data point. For example, today is 2nd March 2019, then I'd need data point from 1st March 2018 to 2nd March 2019.

Please note, once you have the data points for last one-year set, you can update this on a daily basis, which means the daily closing prices are recorded.

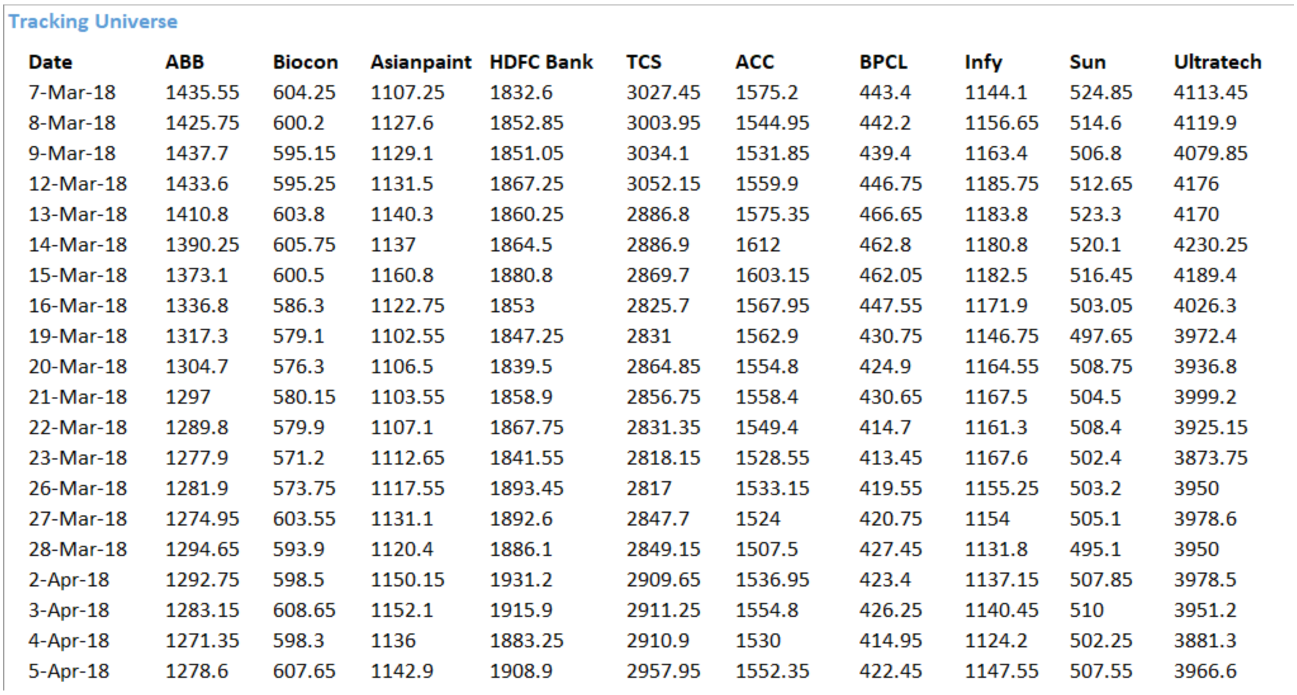
Step 3 – Calculate returns

This is a crucial part of the strategy; in this step, we calculate the returns of all the stocks in the tracking universe. As you may have already guessed, we calculate the return to get a sense of the momentum in each of the stocks.

As we discussed earlier in this chapter, one can calculate the returns on any time frequency, be it daily/weekly/monthly or even yearly returns. We will stick to yearly returns for the sake of this discussion, however, please note; you can add your own twist to the entire strategy and calculate the returns on any time frequency you wish. Instead of yearly, you could calculate the half-yearly, monthly, or even fortnightly returns.

So, at this stage, you should have a tracking universe consisting of about 150-200 stocks. All these stocks should have historical data for at least 1 year. Further, you need to calculate the yearly return for each of these stocks in your tracking universe.

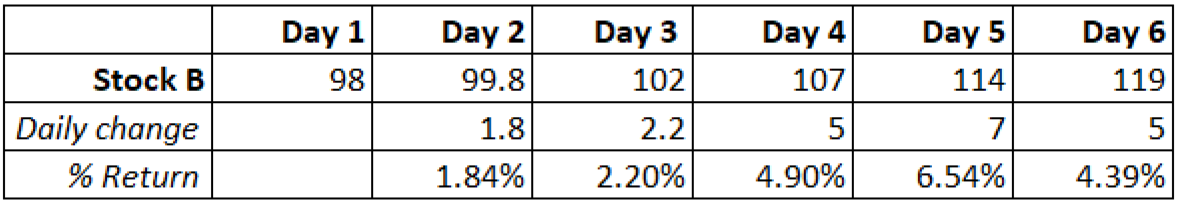
To help you understand this better, I've created a sample tracking universe with just about 10 stocks in it.



Tracking Universe

Date	ABB	Biocon	Asianpaint	HDFC Bank	TCS	ACC	BPCL	Infy	Sun	Ultratech
7-Mar-18	1435.55	604.25	1107.25	1832.6	3027.45	1575.2	443.4	1144.1	524.85	4113.45
8-Mar-18	1425.75	600.2	1127.6	1852.85	3003.95	1544.95	442.2	1156.65	514.6	4119.9
9-Mar-18	1437.7	595.15	1129.1	1851.05	3034.1	1531.85	439.4	1163.4	506.8	4079.85
12-Mar-18	1433.6	595.25	1131.5	1867.25	3052.15	1559.9	446.75	1185.75	512.65	4176
13-Mar-18	1410.8	603.8	1140.3	1860.25	2886.8	1575.35	466.65	1183.8	523.3	4170
14-Mar-18	1390.25	605.75	1137	1864.5	2886.9	1612	462.8	1180.8	520.1	4230.25
15-Mar-18	1373.1	600.5	1160.8	1880.8	2869.7	1603.15	462.05	1182.5	516.45	4189.4
16-Mar-18	1336.8	586.3	1122.75	1853	2825.7	1567.95	447.55	1171.9	503.05	4026.3
19-Mar-18	1317.3	579.1	1102.55	1847.25	2831	1562.9	430.75	1146.75	497.65	3972.4
20-Mar-18	1304.7	576.3	1106.5	1839.5	2864.85	1554.8	424.9	1164.55	508.75	3936.8
21-Mar-18	1297	580.15	1103.55	1858.9	2856.75	1558.4	430.65	1167.5	504.5	3999.2
22-Mar-18	1289.8	579.9	1107.1	1867.75	2831.35	1549.4	414.7	1161.3	508.4	3925.15
23-Mar-18	1277.9	571.2	1112.65	1841.55	2818.15	1528.55	413.45	1167.6	502.4	3873.75
26-Mar-18	1281.9	573.75	1117.55	1893.45	2817	1533.15	419.55	1155.25	503.2	3950
27-Mar-18	1274.95	603.55	1131.1	1892.6	2847.7	1524	420.75	1154	505.1	3978.6
28-Mar-18	1294.65	593.9	1120.4	1886.1	2849.15	1507.5	427.45	1131.8	495.1	3950
2-Apr-18	1292.75	598.5	1150.15	1931.2	2909.65	1536.95	423.4	1137.15	507.85	3978.5
3-Apr-18	1283.15	608.65	1152.1	1915.9	2911.25	1554.8	426.25	1140.45	510	3951.2
4-Apr-18	1271.35	598.3	1136	1883.25	2910.9	1530	414.95	1124.2	502.25	3881.3
5-Apr-18	1278.6	607.65	1142.9	1908.9	2957.95	1552.35	422.45	1147.55	507.55	3966.6

The tracking universe contains the data for the last 365 days. The 1-year returns are calculated as well –



	Day 1	Day 2	Day 3	Day 4	Day 5	Day 6
Stock B	98	99.8	102	107	114	119
Daily change		1.8	2.2	5	7	5
% Return		1.84%	2.20%	4.90%	6.54%	4.39%

If you are wondering how the returns are calculated, then this is quite straight forward, let us take the example of ABB –

Return = [ending value/starting value]-1

= [1244.55/1435.55]-1

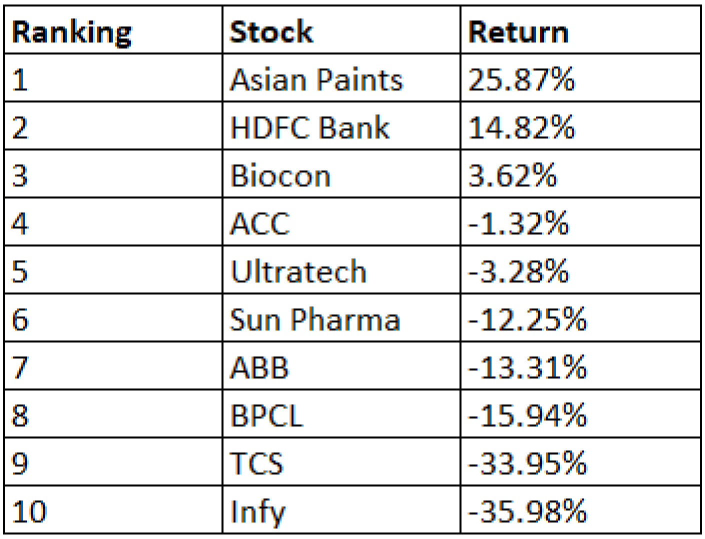
= **-13.31%**

Quite straightforward, I guess.

Step 4 – Rank the returns

Once the returns are calculated, you need to rank the returns from the highest to the lowest returns. For example, Asian paints have generated a return of 25.87%, which is the highest in the list. Hence, the rank of Asian paints is 1. The second highest is HDFC Bank, so that will get the 2nd rank. Infosys's return, on the other hand, is -35.98%, the lowest in the list, hence the rank is 10. So on and so forth.

Here is the 'return ranking' for this portfolio –



Ranking	Stock	Return
1	Asian Paints	25.87%
2	HDFC Bank	14.82%
3	Biocon	3.62%
4	ACC	-1.32%
5	Ultratech	-3.28%
6	Sun Pharma	-12.25%
7	ABB	-13.31%
8	BPCL	-15.94%
9	TCS	-33.95%
10	Infy	-35.98%

If you are wondering why the returns are negative for most of the stocks, well then, that's how stocks behave when deep corrections hit the market. I wish, I had opted to discuss this strategy at a better point

So what does this ranking tell us?

If you think about it, the ranking reorders our tracking universe to give us a list of stocks from the highest return stock to the lowest. For example, from this list, I know that Asian Paints has been the best performer (in terms of returns) over the last 12 months. Likewise, Infy has been the worst.

Step 5 – Create the portfolio

A typical tracking universe will have about 150-200 stocks, and with the help of the previous step, we would have reordered the tracking universe. Now, with the reordered tracking universe, we are good to create a momentum portfolio.

Remember, momentum is the rate of change of return and the return itself is measured on a yearly basis.

A good momentum portfolio contains about 10-12 stocks. I'm personally comfortable with up to 15 stocks in the portfolio, not more than that. For the sake of this discussion, let us assume that we are building a 12 stocks momentum portfolio.

The momentum portfolio is now simply the top 12 stocks in the reordered tracking universe. In other words, we buy all the stocks starting from rank 1 to rank 12. In the example we were dealing with, if I were to build a 5 stock momentum portfolio, then it would contain –

- Asian Paints
- HDFC Bank
- Biocon
- ACC
- Ultratech

The rest of the stocks would not constitute the portfolio but will continue to remain in the tracking universe.

What is the logic of selecting this subset of stocks within the tracking universe, you may ask?

Well, read this carefully – if the stock has done well (in terms of returns generated) for the last 12 months, then it implies that the stock has good momentum for the defined time frame. The expectation is that this momentum will continue onto the 13th month as well, and therefore the stock will continue to generate higher returns. So if you were to buy such stocks, then you are to benefit from the expected momentum in the stock.

Clearly, this is a claim. I do not have data to back this, but I have personally used this exact technique for a couple of years with decent success. It is easy to back-test this strategy, and I encourage you to do so.

Back in the days, my trading partner and I were encouraged to build this momentum portfolio after reading this **'Economist'** article. You need to read this article before implementing this strategy.

Once the momentum portfolio stocks are identified, the idea is to buy all the momentum stocks in equal proportion. So if the capital available is Rs. 200,000/- and there are 12 stocks, then the idea is to buy Rs. 16,666/- worth of each stock (200,000/12).

By doing so, you create an equally weighted momentum portfolio. Of course, you can tweak the weights to create a skewed portfolio, there is no problem with it, but then you need to have a solid reason for doing so. This reason should come from backtested results.

If you like to experiment with skewed portfolios, here are few ideas –

- 50% of capital allocation across the top 5 momentum stocks (rank 1 to 5), and 50% across the remaining 7 stocks
- Top 3 stocks get 40% and the balance 60% across 9 stocks
- If you are a contrarian and expect the lower rank stocks to perform better than the higher rank stocks, then allocate more to last 5 stocks

So on and so forth. Ideally, the approach to capital allocation should come from your backtesting process, this also means you will have to backtest various capital allocation techniques to figure out which works well for you.

Step 6 – Rebalance the portfolio

So far, we have created a tracking universe, calculated the 12-month returns, ranked the stocks in terms of the 12-month returns, and created a momentum portfolio by buying the top 12 stocks. The momentum portfolio was built based on the 12-month performance, with a hope that it will continue to showcase the same performance for the 13th month.

There are few assumptions here –

- The portfolio is created and bought on the 1st trading day of the month
- The above implies that all the number crunching happens on the last day of the month, post-market close
- Once the portfolio is created and bought, you hold on to the stocks till the last day of the month

Now the question is, what really happens at the end of the month?

At the end of the month, you re-run the ranking engine and figure out the top 10 or 12 stocks which have performed well over the last 12 months. Do note, at any point we consider the latest 12 months of data.

So, we now buy the stocks from rank 1 to 12, just like the way we did in the previous month. From my experience, chances are that out of the initial portfolio, only a hand full of stocks would have changed positions. So based on the list, you sell the stocks which no longer belongs in the portfolio and buy the new stocks which have featured in the latest momentum portfolio. In essence, you rebalance the portfolio and you do this at the end of every month.

So on and so forth.

16.4 – Momentum Portfolio variations

Before we close this chapter (and this module), I'd like to touch upon a few variations to this strategy.

The returns have been calculated on a 12-month portfolio and the stocks are held for a month. However, you don't have to stick to this. You can try out various options, like –

- Calculate return and rank the stocks based on their monthly performance and hold the portfolio for the month
- Calculate return and rank the stocks based on fortnightly performance and hold the portfolio for 15 days
- Rank on a weekly basis and hold for a week
- Calculate on a daily basis and even do an intraday momentum portfolio

As you can see, the options are plenty and it's only restricted by your imagination. If you think about what we have discussed so far, the momentum portfolio is price based. However, you can build a fundamental based momentum strategy as well. Here are a few ideas –

- Build a tracking universe of fundamentally good stocks
- Note the difference in quarterly sales number (% wise)
- Rank the stocks based on quarterly sales. Company with the highest jump in sales gets rank one and so on
- Buy the top 10 – 12 stocks
- Rebalance at the end of the quarter

You can do this on any fundamental parameter – EPS growth, profit margin, EBITDA margin etc. The beauty of these strategies is that the data is available, hence backtesting gets a lot easier.

16.5 – Word of caution

As good as it may seem, the price based momentum strategy works well only when the market is trending up. When the markets turn choppy, the momentum strategy performs poorly, and when the markets go down, the momentum portfolio bleeds heavier than the markets itself.

Understanding the strategy's behaviour with respect to market cycle is quite crucial to the eventual success of this portfolio. I learned it the hard way. I had a great run with this strategy in 2009 and '10 but took a bad hit in 2011. So before you execute this strategy, do your homework (backtesting) right.

Having said all of that let me reassure you – a price based momentum strategy, if implemented in the right market cycle can give you great returns, in fact, better more often than not, better than the market returns.

Good luck and happy trading.

Key takeaways from this chapter

- Momentum is the rate of change of return and can be measured across any time frame
- A price based momentum portfolio consists of stocks which have exhibited highest momentum over the desired time frame
- Tracking universe should be carefully populated. BSE 500 is a good tracking universe
- Calculate the returns for the tracking universe
- Rank the stocks based on highest to lowest return
- The momentum portfolio is simply the top 12 or 15 stocks
- The expectation is that the momentum will continue during the holding period

- The asset allocation technique can vary based on backtesting Equally weighted portfolio is a good asset allocation technique
- Momentum can be measured on fundamental data as well – growth in sales, EBITDA margins, EPS growth, net profit margin etc.
- Price based momentum works best in an upward trending market and not really in a sideways or a down trending market.