Statistical Arbitrage on Banking stocks using Linear Regression

Contents

1 Introduction

2 Data Acquisition

3 EDA

• Banking stock close prices

• Normalized plot of banking stocks

• Correlation plot

• Correlation plot highlighting pairs with strong relationship

• Pair plot

4 Idea

• Stationarity

• Cointegration

• Trading opportunity

5 Modeling

5 Modeling

• Data preprocessing

• Training model

• Calculating spreads

• Running mean reversion backtest

• Results

Introduction

In financial markets and in the real world there are time series which are cointegrated. In general terms if two time series are ‘co integrated’ then it means, that the two time-series 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. This type of time series is exactly what we want to find in financial markets.

So, in this project I will try and find those cointegrated relationships in banking stocks

So, to try to find a cointegrated time series we use linear regression. So, what linear regression does in basic terms comes with the relationship between Independent variable X and dependent variable Y. That relationship is described in terms of slope. so, slope captures how much Y would change with change in X.

Once we get a relationship then we can use that relationship to predict the value of Y from X. But more often linear regression has a hard time predicting dependent variables properly. so, we are not going to use linear regression predictions to directly make trades.

Instead of focusing on predictions from the linear regression model we are going to focus on errors made by the model also known as residuals. so, we will check whether the residuals of a model are stationary or not meaning the residuals mean and variance is in tight range and autocorrelation of residual is close to 0. This means a residual time series is stationary.

Stationarity of residuals confirms that two stocks are cointegrated. they move closely together and if at all some deviation happens that's temporary, they will converge back. That’s what a cointegrated series is.

Data

a. I am Using NSE banking stocks and bank nifty index cash 1min data from 2014 to 2020

b. Stock data is a time series data with Open, High, Low, Close and volume columns c. I am only taking banking stocks data for this project but the same approach can be applied to the stocks from different sector

d. Data is relatively clean

e. Open column has the opening price for the stock for the given timestamp f. High column has the high price stock went in that time frame

g. Close column has the close price of the stock for that time frame

h. Volume has the amount of trades happened in the given time frame

Below is the list of stocks and bank nifty index cash data that i am using for this project:

['BANKNIFTY', 'AXISBANK', 'BANKBARODA,

'BANKINDIA','HDFC', 'HDFCBANK.csv',

'ICICIBANK', 'IDBI', 'KOTAKBANK.csv',

'PNB', 'SBIN']

data acquired from my current firm's data vendor data cannot be shared due copyright issues

Snapshot of 1min OHLC HDFC stock data

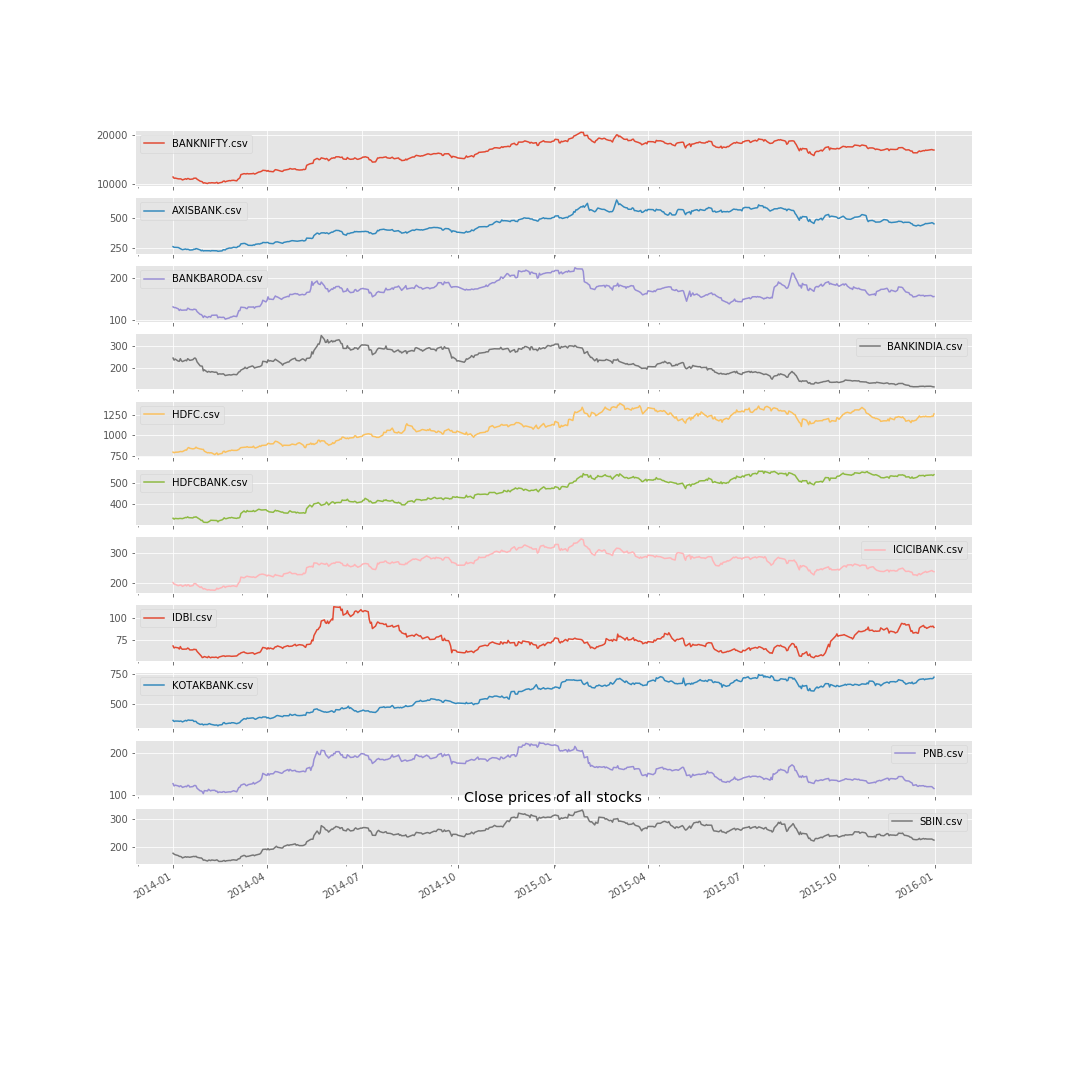


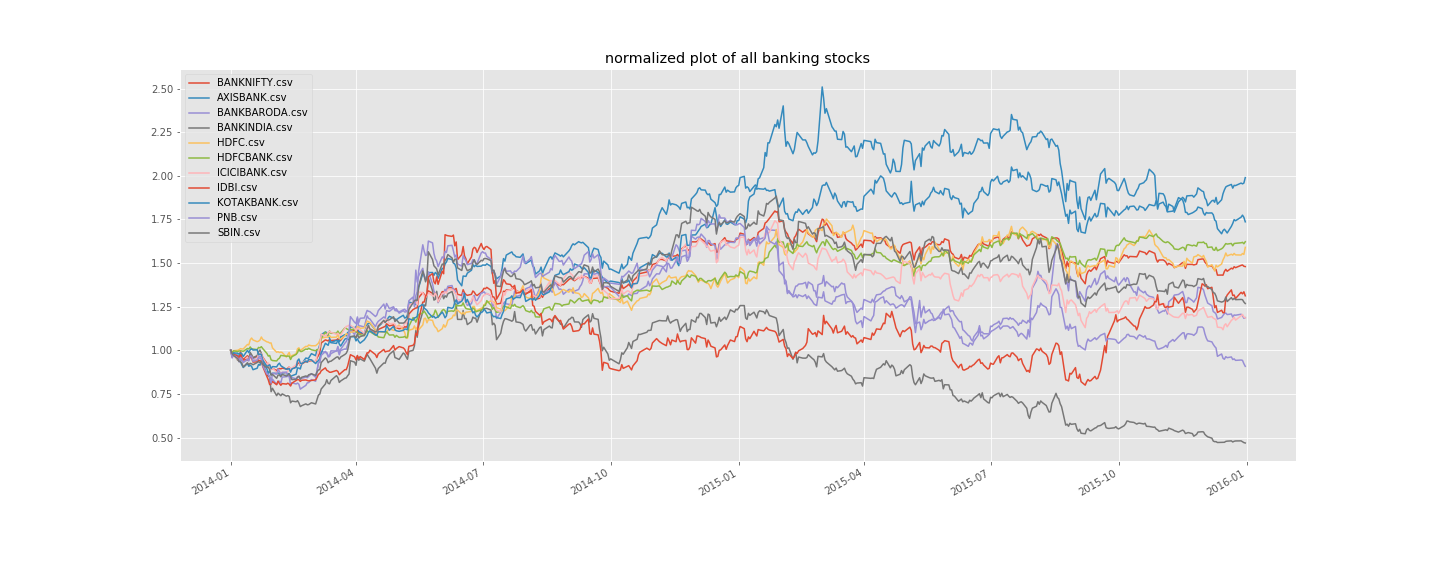
EDA

In EDA I am keen on getting to know which stocks have high negative or positive Correlation. So, I can decide which two stock timeseries to find cointegration in. This narrows down my search. But it is not necessary that correlated pairs will be cointegrated as well but there is more chance of pairs being cointegrated if they have strong correlation as well. So that's why I will use correlation to filter down stocks which have higher negative or positive correlation out of numerous possible banking pairs combinations.

I am doing EDA on data from 2014 - 2015 as we want to avoid look-ahead bias while selecting pairs . as I would be testing my hypothesis on data from 2016 - 2020.

**Close prices of all banking stocks**

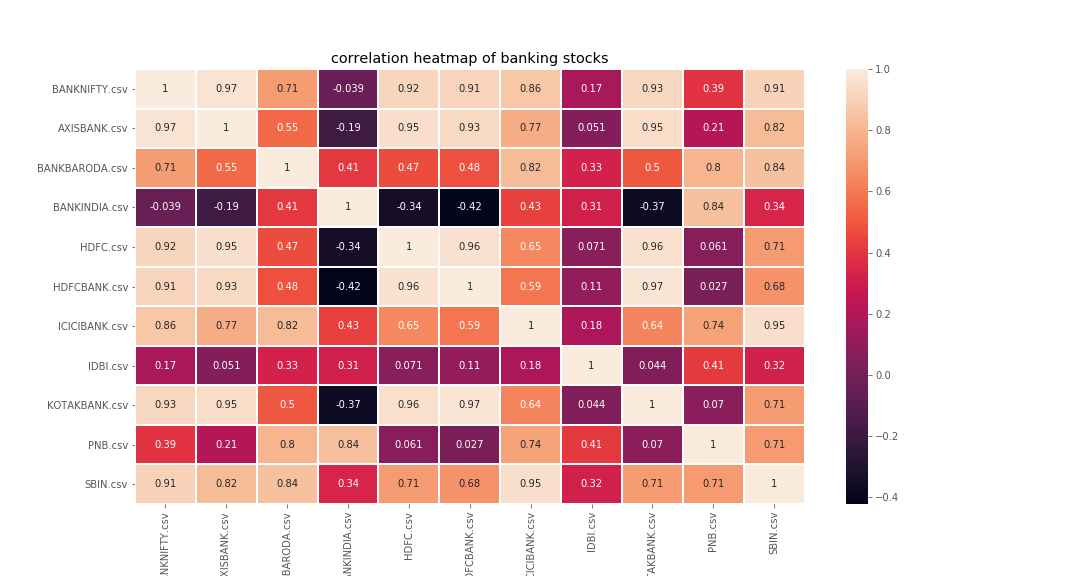


**Normalized Plot of banking stocks to get an idea of the performance** 

**Pair plot**



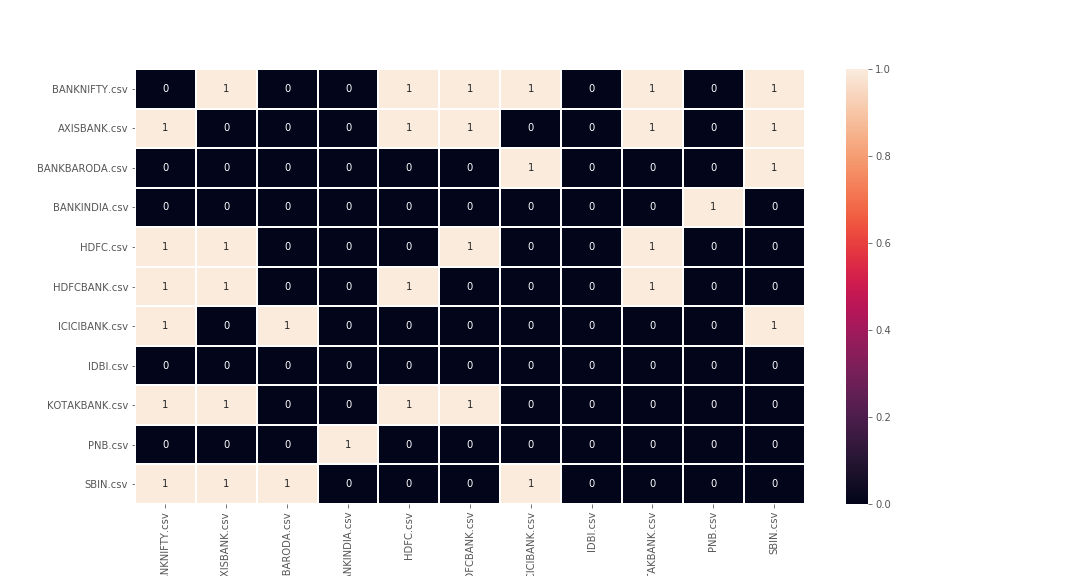
**Correlation heat map**



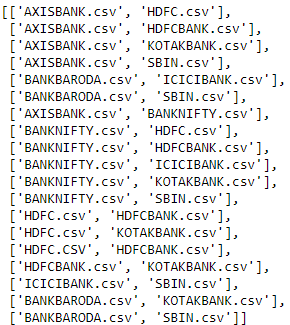
**Pairs with strong positive or negative correlation heatmap**

• 1 if (correlation > 0.8) or (correlation < -0.8)

• 0 for correlation between 0.8 and –0.8



So below are the list of pair i have selected from correlation matrix to run further research and look for trading opportunities in these pairs as they have strong neg or positive correlation

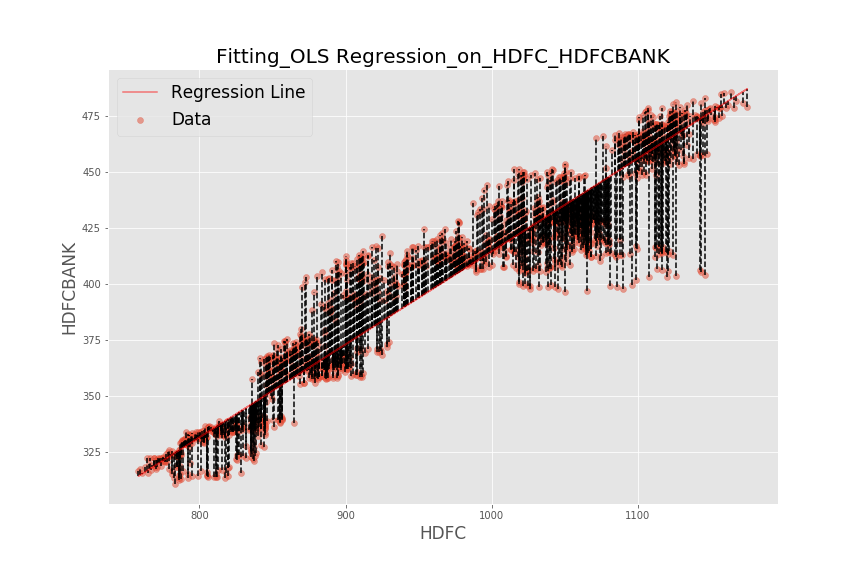


**4 Idea**

• Cointegration

• Stationarity

• Mean reversion trading opportunity



**Cointegration**

So, the idea revolves around finding some stock Pairs which are cointegrated. So, to reiterate what cointegration is that the two time-series 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.

**But 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 we calculate a spread using (actual y - predicted y) spread is nothing but residuals from the regression model. the entire play is in the residuals.

Residuals is itself a time series whether a pair is cointegrated or not can accessed by the stationarity of the residuals

**Stationarity**

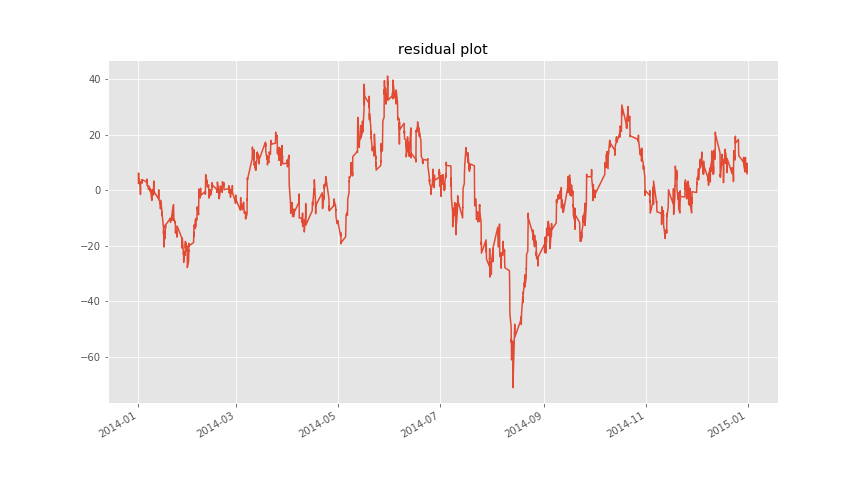
A time series is considered stationary if it has attributes listed below.

**Assumptions of a stationary time series:**

• 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 very less autocorrelation in the time series in simple words, say value n should not be correlated with value before it. which states that time series is not trending



so above in picture we have residuals of hdfc which is independent variable and hdfc bank which is dependent variable residuals if we look at series what do u think series is stationary or trending by eyeballing we can say its mean reverting series. I have confirmed it by running a statistical test as well that i have discussed below

**ADF test**

To check whether the series is stationary or not we run an ADF test. ADF test automatically checks the conditions stated above.we get t-stat value as output from an ADF test and if t-stat value is lower than the critical value we can conclude that the series is stationary.

Null and alternative hypothesis in ADF test

* H0 series is trending
* H1 series is stationary

critical values for ADF test are:

* 1%: -3.4342184244873657 (99% percent confidence we reject H0)
* 5%: -2.8632486915862247 (95% percent confidence we reject H0)
* 10%: -2.567679662838086 (90% percent confidence we reject H0)

**ADF test t-stat value for above HDFC and HDFC BANK residuals time series is : -3.0176525849531557**

**So -3.01 is lower than -2.86 so with 95% confidence we can say series is stationary**

One thing to keep in mind: Residuals from a stock pair might not always be stationary in the real world. More often than not we will find stationarity in patches.

* Meaning pairs which are cointegrated today might lose cointegration after some time.
* So, for that we will use rolling ADF t-stat values. so we can access at every data point whether the past n values of residuals is stationary or not.

**Now the question how we can benefit from a stationary residual time series**

Well **y = Slope \* x + error** means that price of dependent stock Y is equal to price of independent stock X multiplied by slope + some error.

Now let's understand residuals more intuitively so residual is basically (y – y-pred ). And y-pred is X multiplied by slope.

* if we subtract (y – y-pred) and we get a positive residual that means the actual value is greater than predicted.
* And if the residual is negative this means actual value is smaller than predicted.

As there is no asset as residuals trading on exchange. So, if we want to trade residuals like stocks two things can be done, we can set up a position in which we are long the spread or set up a position where we are short the spread.

• If we want buy spread then we buy y stock and sell (X \* slope) qty of X stock.

• And if we want to short spread then we sell y stock and buy (X \* slope) qty of X stock.

**Trading opportunity**

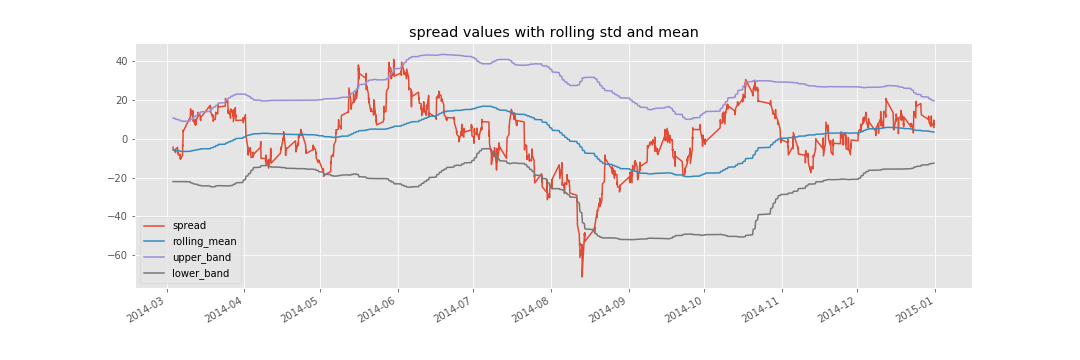
so, remember the equation **`y = slope \* X + error`**. So the spread is an error. And that is what we are trying to capture. So, when residuals go above +2std or below –2std trading opportunity arises and as we know we select pairs whose residual are stationary then the properties of normal distribution can be applied.

• Before checking residual positions we have to make sure that residuals are stationary. Meaning are ADF test t-stat values should be below the critical values that we decide based on the confidence we want for stationarity of the series ie if we want 95% confidence that the series is stationary, we go with -2.8632486915862247 as critical value.

• When the spread goes below - 2std there is an opportunity to buy the spread. As spread is not traded on exchange so to trade it. So, we buy Y stock and sell (X \* slope) qty of X stock.

• When the spread goes +2 std there is an opportunity to sell spread. So, we sell Y stock and buy (X \* slope) qty of X stock.

• Target should be mean when they converge



**Work flow of stat arb on one pair (x = AXISBANK,y = HDFC)**

**• Data preprocessing**

**• Train model**

**• Calculating spreads**

**• Calculating rolling ADF test t-stat values**

**• Running mean reversion back test**

**• Results**

**Data preprocessing**

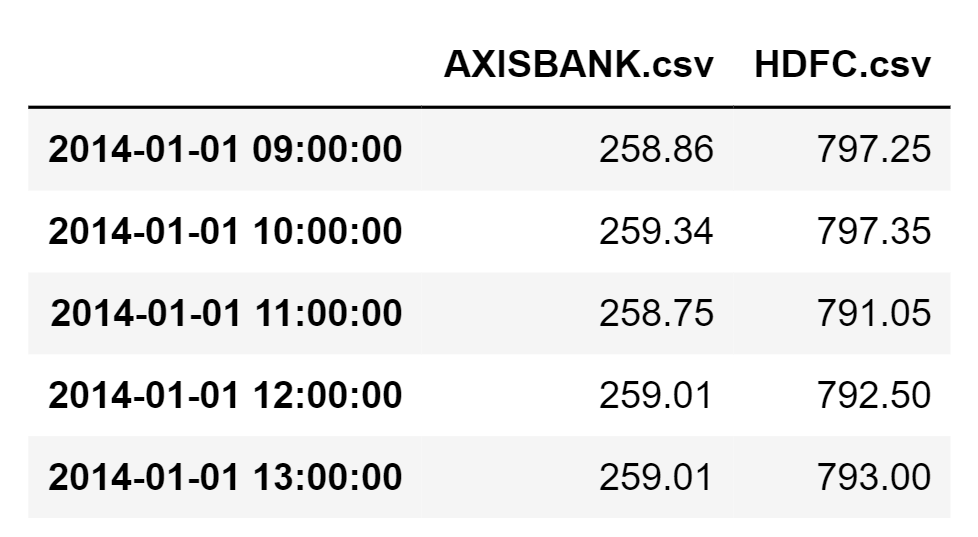
• First, load data get x and y

• Keeping only Close column data of x and y

• Concatenate x and y in one data frame

• Resample the data frame to desired frequency

• Drop nan values if generated while concatenating data frame

In below snapshot we resampled data from 1 min to 60 mins

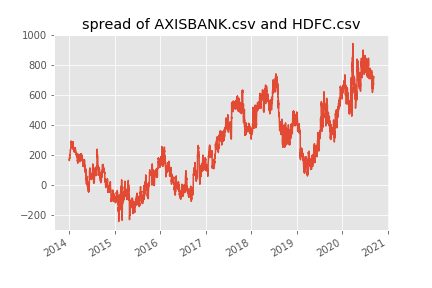
**Training linear regression model**

* Get train data set ready wich in the this case is : 2014 - 2015
* Decide Independent and dependent variable (x = AXISBANK, y = HDFC)
* And the run linear regression on it and get the slope coefficient

**Slope coefficient with (x = AXISBANK, y = HDFC) is 2.4243012726388877**

**Calculate the Spread using Slope coefficient**

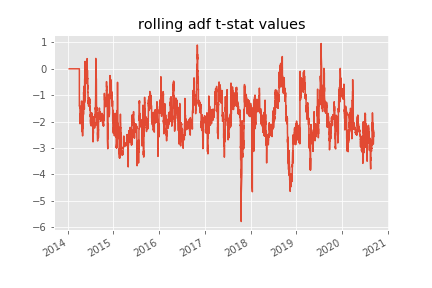
* calculate spread (actual – Predicted)
* calculating spread for the full data 2014 - 2020



**Calculating rolling ADF t-stat values with n = 400**

• getting rolling ADF t-stat values of the spread let us now if past n values series acted stationary or not.

• Idea is for past n values, if the series acted stationary then only, we will go ahead and look for trade.

• if not, we will not look to trade until series again starts showing attributes of stationarity 

**Calculating rolling mean, rolling std,upper band and lower band on Spread**

first we will also calculate rolling mean and rolling std with rolling period = 60 for this example

* then calculate upper band which is rolling mean +2std which is nothing but (rolling mean + (rolling std \* 2))
* calculate lower band which is (rolling mean - (rolling std \* 2)). This helps us know whether spread has deviated too much from mean.

**Mean reversion strategy back test on AXIS BANK and HDFC from 2016 - 2020**

So, trading rules for a back test are as follows:

Look to trade if ADF t-stat values are below critical value.

strategy rules for short

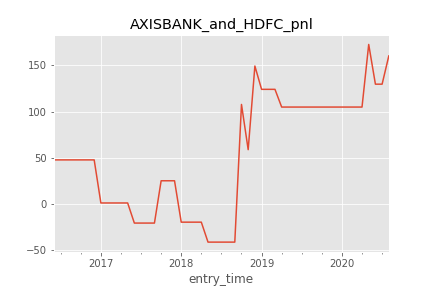
• if the spread goes `above +2std` and is `below +2.5std` take short trade

• And once short trade is taken `target is mean` and `stop is +3 std`

strategy rules for long

• if the spread is `below -2std` and `greater than -2.5std` we take a long trade

• once the long trade is taken `target is mean` and `stop is -3 std`



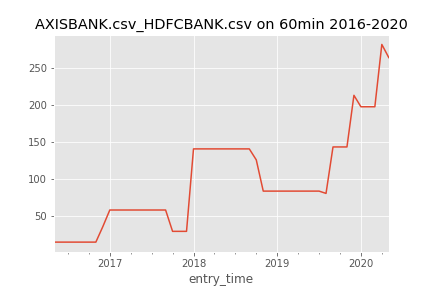
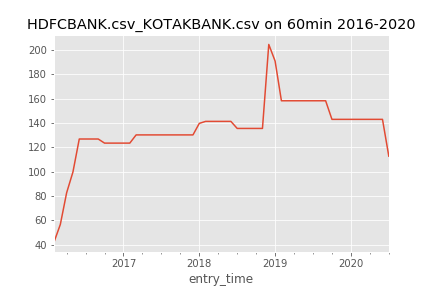
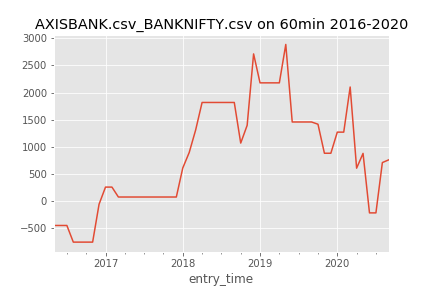
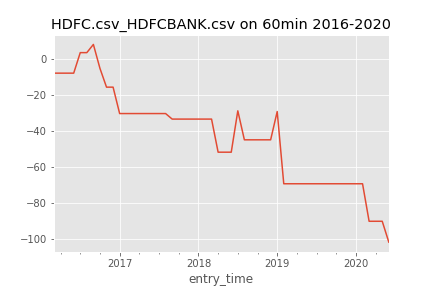
above is the p&l of axis bank and hdfc bank from 2016 - 2020 that we got from full statistical arbitrage workflow now will scale this workflow to multiplepairs

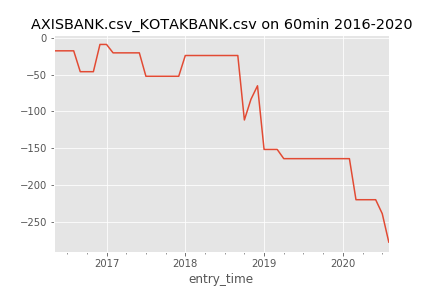
**Scaling statistical arbitrage backtest on multiple pairs on 60 min time frame from 2016 - 2020**

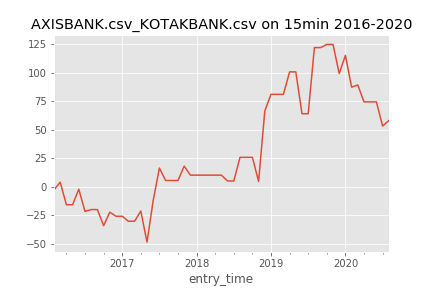


**So above pnl is in absolute points**

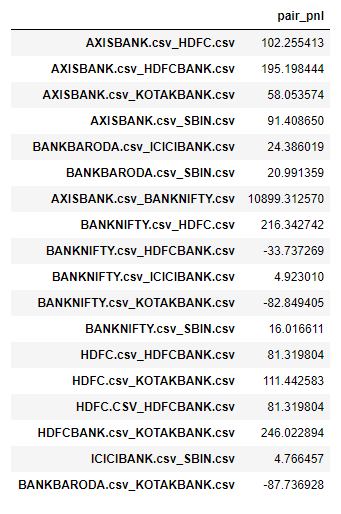
**Lets plot cumulative p&l of some pairs**



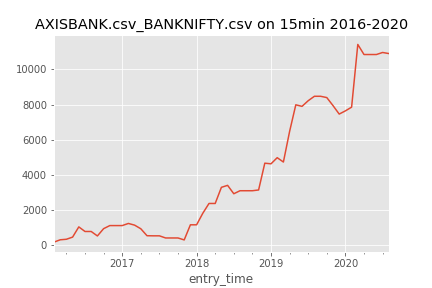
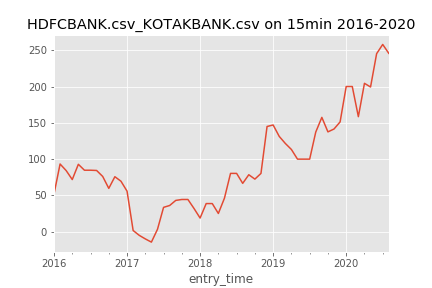
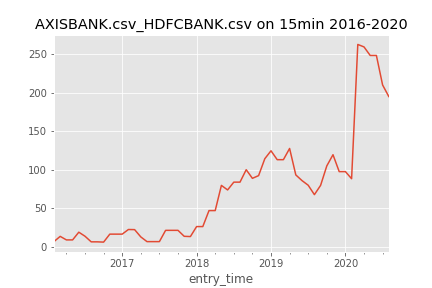
****

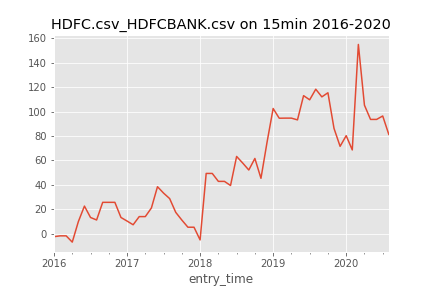
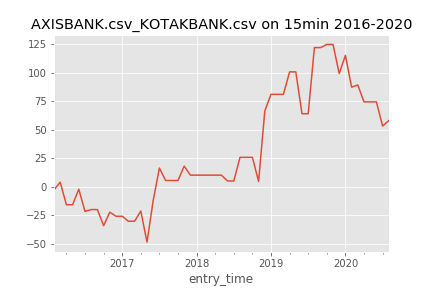


**Scaling statistical arbitrage backtest on 15 min on multiple pairs-results**

****

**Lets plot cumulative p&l of some pairs**

****

****

**Conclusion**

**statistical arbitrage does work if done properly as we can see in the cumulative p&l of individual pairs key is to selecting pairs and using right parameters like adf critical values threshold, std , rolling n in rolling calculations etc so that's there . that will do as a part of further research**