**Estrutura dos Dados**

1. Em cada arquivo, a primeira linha é a identificação do arquivo e a ultima linha é o rodapé do que apresenta o total de linhas do arquivo. Estas duas linhas (cabeçalho e rodapé) precisam ser excluídas, caso contrário da erro na hora de juntar os arquivos.
2. Excluídas a linha de cabeçalho e rodapé, vou te passar o que é cada coluna. No arquivo NEG\_LAYOUT\_portuguese da pra saber certinho em qual posição começa cada coluna, que são delimitadas por ';'. As marcadas em amarelo são as necessárias para o cálculo das variáveis. As demais podem ser excluídas. Uma nova coluna precisa ser criada, será marcada em vermelho.

Coluna 1: Data

Coluna 2: Ticker/identificação do ativo.

Coluna 3: Número do negócio

Coluna 4: Preço (Os valores são separados por ponto, então quando for 1000.5, quer dizer mil reais e cinquenta centavos.)

Coluna 5: Quantidade negociada

Coluna 6: Horário da negociação (formato HH:MM:SS.NNN)

Coluna 7: Indicador de Anulação

Coluna 8: Data da oferta de compra

Coluna 9: Número sequencial da oferta de compra

Coluna 10: Número de geração (GenerationID) da Oferta de compra

Coluna 11: Código que identifica a condição da oferta de compra (Se for 1 é Compra e se for 2 é Venda)

Coluna 12: Data da oferta de venda

Coluna 13: Número sequencial da oferta de venda

Coluna 14: Número de geração (GenerationID) da Oferta de venda

Coluna 15: Código que identifica a condição da oferta de venda

Coluna 16: Código que identifica se o negócio direto foi intencional

Coluna 17: Código de identificação da corretora de compra

Coluna 18: Código de identificação da corretora de venda

Coluna a ser criada – Volume Negociado: Coluna 4 (preço) x Coluna 5 (quantidade negociada). Assim calcularemos o volume de negociação.

1. A coluna 1 data e coluna 6 horário poderia ser “Juntadas” em uma só, porque quando juntar os arquivos todos em um único data frame, é preciso que o arquivo esteja na ordem da data e na ordem cronológica certinha.
2. Feito isso, eu preciso poder filtrar, desse data frame criado até aqui, somente os tickers que eu quero. Por exemplo, desse data frame, vamos calcular as variáves só do Dólar Futuro, então é preciso passar pelo data frame original e selecionar as linhas que tenham no ticker as letras DOL e copiá-las pra outro data frame. O ticker tem que ter 6 dígitos. Tem uns que tem mais, mas são linhas que precisam ser ignoradas, pois são testes da bolsa. Não são negociações reais. AS reais são assim por exemplo: DOLG19. Eu preciso poder escolher pra qual ativo eu quero fazer os cálculos. Assim, quando eu quiser é só trocar o código do ticker e calcular de novo que da pra escrever outro artigo. O mais urgente é o dólar comercial (DOL) e o DI (DI1). Os ativos precisam ser separados pra calcular as variáveis porque vai ter uma regressão pra cada ativo. PS: Pretendo rodar a regressão em outro softwere.

**Cálculo das Variáveis**

Feito o data frame com as colunas escolhidas no passo 2, na ordem cronológica e do ativo filtrado no passo 04, pode-se enfim calcular as variáveis.

**1. VPIN**

A primeira variável é a VPIN, cujo cálculo se dá pela fórmula:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

Em que:

é o volume de ordens de venda de cada *bucket*;

é o volume de ordens de compra de cada *bucket*;

*V* é o volume de cada *bucket* e

*n* é o número de *buckets* utilizados.

Os *buckets* são lotes equivalentes a 1/50 avos do volume negociado por dia. Por exemplo, foi criada uma nova coluna (Preço x Quantidade negociada) que é o volume negociado. É preciso somar o volume negociado dia a dia e criar o bucket = volume negociado no dia/50, e desse volume do bucket, verificar, pela coluna 11, o que foi compra (1) e o que foi venda (2). Por exemplo:

Verificou-se que o volume de dólar (DOL) negociado no dia 02/01/2019 foi de 100 mil reais. Então cada bucket (50 no dia) será de 2.000,00 (100.000/50). Esse valor será o mesmo para o dia 02/01/2019 e só vai mudar no dia 03/01/2019, quando o volume será diferente. Bem, pra calcular a VPIN, deve-se olhar cada bucket do dia e ver qual volume de compra e qual volume de venda e assim fazer o cálculo. Vamos supor que o primeiro bucket do dia 02/01/2019, dos 2000,00, 1500,00 seja volume de compra e 500,00 seja de venda. Então ficaria:

VPIN = |500-1500| / 50x2000

VPIN = 1000/100.000 = 0,01

Nesse sentido, como será usado 50 buckets diários, também terá 50 observações de VPIN por dia. Se um mês tiver 20 dias de negociações, o número de observações da VPIN será de 20\*50 = 1000.

**2. Bid-Ask-Spread**

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

Em que:

é a diferença entre a oferta de venda e a oferta de compra no tempo *t*;

é a oferta de venda no tempo *t*; e

é a oferta de compra no tempo *t*.

Para gerar a regressão, eu preciso do mesmo número de observações em todas as variáveis. Dessa forma, como a VPIN terá 50 observações por dia, preciso que o bid-ask-spread também tenha 50 observações. Assim ele deve ser calculado por bucket, o mesmo criado pra calcular a VPIN, só que, ao invés de olharmos o volume, olharemos o preço. O bid-ask-spread será calculado subtraindo o preço da ultima oferta de venda, pela primeira oferta de compra dentro de cada bucket. Então, se o preço a primeira compra (1 na coluna 11) de dólar foi de 3000,00 as 09:00:01 e a última oferta de venda (2 na coluna 11) foi de 3001,00 as 10:00:00. O bid-ask-spread será de 3001 – 3000 = 1,00.

**3. Número de Transações**

O número de transações se encontra na coluna 5: Quantidade negociada. Preciso do número de transações em cada bucket, então é só somar a quantidade negociada por bucket, o mesmo que foi feito para o cálculo da VPIN, mas agora preciso saber apenas o número de contratos negociados dentro do bucket. Assim, também terei 50 observações por dia.

**4. Volatilidade dos Retornos**

Dentro de cada bucket utilizado na VPIN, calcular o retorno linha por linha [(P t+1)/Pt] -1 (Assim fica em decimal). A partir desse retorno, calcula a volatilidade (variância) por bucket. Tendo assim, 50 observações por dia.

**5. Volatilidade dos Preços**

Só calcular a variância dos preços dentro de cada bucket, tendo também 50 observações. O item 4 e 5 devem ficar parecidos, vou escolher qual fica melhor na regressão.

**Exportar as variáveis para Excel, gerar gráficos de linha das variáveis e calcular as estatísticas descritivas de cada uma. O arquivo excel pode ser um pro dólar e outro pro di e cada coluna com uma variável calculada.**