Séries Temporais aplicadas ao Mercado Financeiro

Modelo preditivo de preços das ações da B3

22/10/2020

Pós-Graduação em Análise de Big Data

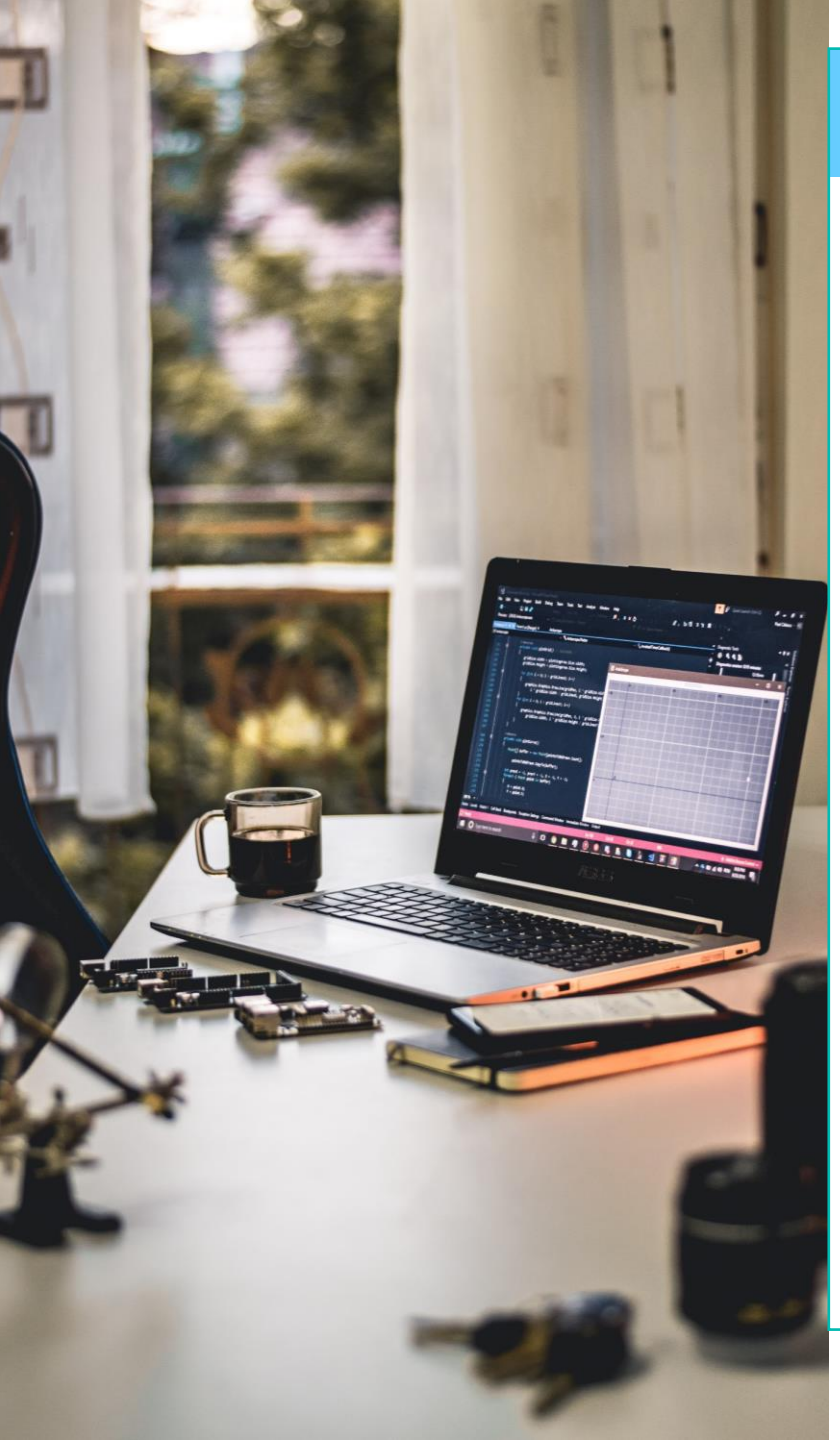
Aluno:
Caíque Tadeu Filipini

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Prof. Dr. Adolpho Walter Pimazoni Canton

Agenda

1. Objetivo
2. Contextualização
3. Os Dados
4. Análise Exploratória
5. Tratamento da Base
6. Modelagem Tradicional
7. Modelagem Avançada
8. Conclusões

1. Objetivo



1. Objetivo

Desenvolver um modelo que faça a **predição dos preços das ações das empresas listadas na B3** e, com isso, recomendar uma carteira de ativos que supere os resultados da Selic do Ibovespa.

Para isso, serão utilizadas as bases de dados históricos (**séries temporais**) dos preços das ações das empresas. Também é possível que sejam utilizados dados históricos de alguns indicadores macroeconômicos contidos no site do Banco Central do Brasil (IPCA, Selic, dólar, etc.), da perfomance do Ibovespa, dos dados fundamentalistas das empresas (lucro, endividamento, ROE, etc.) e de distribuições de dividendos e JCP (juros sobre capital próprio).

Para extração dos dados, serão utilizadas técnicas de **Web Scrapping** e requisições por meio de **API** (quando disponibilizada pela fonte).

Para criar o modelo preditivo, serão aplicadas, inicialmente, as **técnicas tradicionais** de séries temporais (ARIMA, SARIMA, etc.), e posteriormente o problema será abordado com **modelos mais robustos e sofisticados** para tratamento de padrões sequenciais, como:

- Discretização da série temporal para aplicação de algoritmos de Machine Learning; e
- Deep Learning (RNN, LSTM, etc.)

2. Contextualização

2. Contextualização

O **mercado financeiro**, conhecido por sua **volatilidade**, é fortemente influenciado por diversos fatores, vários deles qualitativos e aleatórios, o que dificulta muito a previsão dos preços das ações. Qualquer movimentação política inesperada, por exemplo, poderia causar uma queda brusca generalizada nos ativos da bolsa. Por outro lado, uma notícia de que uma vacina para uma doença crítica teria sido desenvolvida, fortaleceria a confiança dos investidores na economia, implicando diretamente em uma valorização do mercado financeiro. Além do ambiente interno, este setor também é muito **sensível aos indicadores estrangeiros**, principalmente de países como os EUA. Por isso, prever o preço das ações das empresas torna-se uma **tarefa extremamente complexa e desafiadora**.

Contudo, transferindo para os algoritmos apenas a tarefa de encontrar padrões e tendências nos indicadores gerais do mercado financeiro, o ser humano poderia investir seu tempo avaliando outras variáveis que a máquina não consegue capturar. Desta forma, as saídas preditivas do modelo funcionariam não como soberanas e decisivas, mas sim como adicionais às análises do ser humano, **contribuindo e agregando valor na tarefa de compra e venda das ações**.

3. Os Dados

3. Os Dados / 3.1. Base Original

Para o desenvolvimento deste projeto, é necessário um **período significativo** de dados. Contudo, **datas exageradamente distantes possuem pouca influência no cenário atual**. E a influência que possuem, estão representadas, em grande parte, em dados relativamente mais recentes. Sendo assim, a fim de se obter volume suficiente para treinar, validar e testar os modelos, serão extraídos dados **a partir de 01/01/2010**.

A base de dados principal é composta pelas colunas:

- **Date** - Data da negociação
- **Open** - Preço de abertura
- **High** - Maior preço atingido
- **Low** - Menor preço atingido
- **Close** - Preço de fechamento
- **Volume** - Volume negociado
- **Ticker** - Código da empresa

Quantidade de ações listadas na B3:	328
Quantidade de dias no período considerado:	2.569
Quantidade máxima* de registros:	**842.632**

*Como é possível que nem todas as ações listadas na bolsa de valores possuam capital aberto desde o início do período considerado, esta é a quantidade máxima possível de registros que a base geral pode ter, mas não necessariamente a quantidade real.

3. Os Dados / 3.2. Amostra dos Dados

Exemplos da série temporal de cotações históricas do **BBAS3 (Banco do Brasil SA).**

Date	High	Low	Open	Close	Volume	Adj Close	Ticker
03/01/2005	11,15	10,83	10,84	11,00	1.347.300	6,05	BBAS3
04/01/2005	11,09	10,73	11,00	10,73	2.973.600	5,91	BBAS3
05/01/2005	10,77	10,00	10,77	10,57	2.292.300	5,81	BBAS3
06/01/2005	10,57	10,02	10,56	10,43	1.817.100	5,74	BBAS3
07/01/2005	10,65	10,25	10,43	10,33	680.400	5,69	BBAS3
10/01/2005	10,34	10,02	10,34	10,07	1.463.400	5,54	BBAS3
11/01/2005	10,23	9,95	10,13	10,10	1.485.000	5,56	BBAS3
			...				
28/07/2020	35,78	34,55	34,70	35,03	17.055.500	35,03	BBAS3
29/07/2020	35,95	35,15	35,38	35,95	13.101.300	35,95	BBAS3
30/07/2020	35,54	34,52	35,12	34,80	23.154.800	34,80	BBAS3
31/07/2020	34,97	33,57	34,90	33,58	20.701.300	33,58	BBAS3
03/08/2020	35,04	33,78	34,24	34,35	24.030.700	34,35	BBAS3
04/08/2020	34,60	33,01	33,88	33,30	20.774.400	33,30	BBAS3
05/08/2020	33,97	32,90	33,81	33,33	14.184.600	33,33	BBAS3
06/08/2020	34,43	33,14	33,76	34,35	18.267.300	34,35	BBAS3
07/08/2020	34,94	33,60	34,00	34,11	18.561.800	34,11	BBAS3
10/08/2020	34,60	33,75	34,12	34,43	9.726.000	34,43	BBAS3

3. Os Dados / 3.3. Observações

Observações:

1. A base descrita anteriormente é a **base principal**. Contudo, ao longo do desenvolvimento do trabalho, é possível que mais bases de dados sejam extraídas a fim de **agregar valor** ao projeto, como por exemplo:

- dados históricos fundamentalistas das empresas
- dados históricos de distribuição de dividendos e JCP
- indicadores macroeconômicos históricos (ex: Selic, IPCA, dólar, etc.)

2. Também é possível, para efeito de simplificação do trabalho e possibilidade de **processamento e armazenamento** em computador tradicional, que se limite a quantidade de ações analisadas de acordo com algum critério.

3. Os Dados / 3.4. Filtros e Tratamentos

Filtros:

Antes de abordar o problema com todas as ações da B3, é necessário que verifiquemos primeiro se há indícios de padrões presentes nas cotações dos papéis ao longo do tempo. Portanto, a princípio, para facilitar, limitaremos a quantidade de papéis a serem analisados. Se concluirmos que há padrões significativos, podemos extrapolar as análises para as demais empresas. Sendo assim, por motivo de gosto particular, escolheu-se analisar as empresas do setor financeiro, isto é, bancos.

Tratamentos:

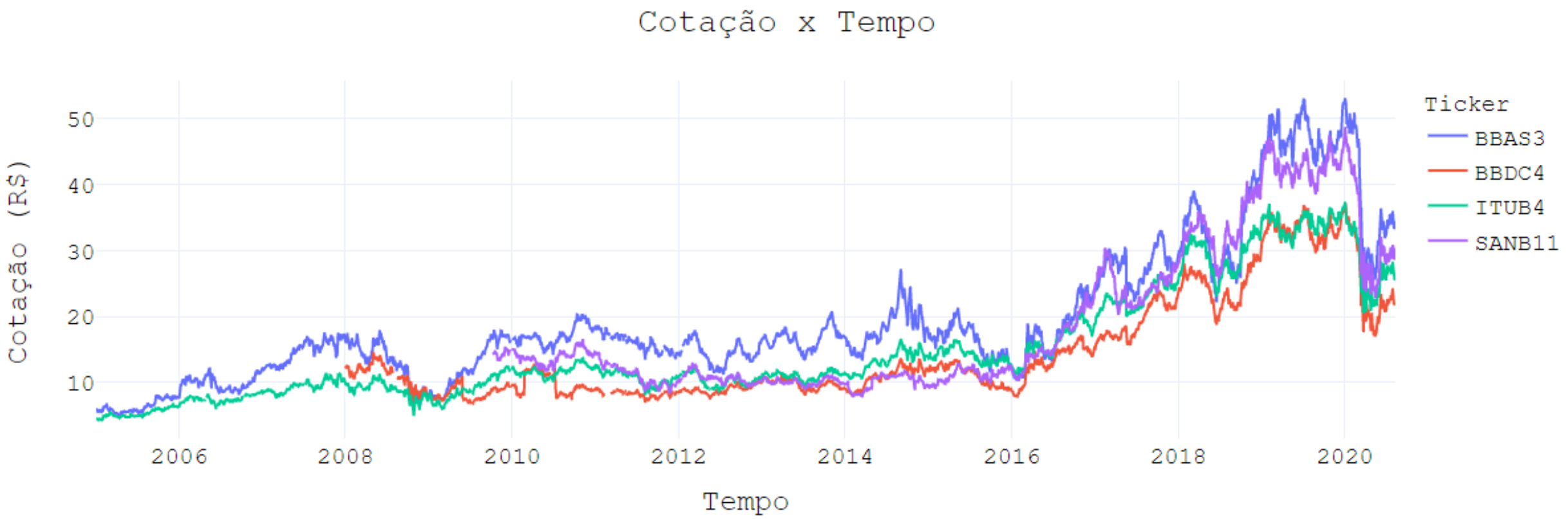
O dado temporal que iremos analisar será o valor ajustado de fechamento das ações, ou seja, a coluna 'Adj Close'. Portanto, é necessário arrumar o layout do DataFrame para que possamos comparar as empresas.

Removemos papéis de empresas iguais. Exemplo: Tanto ITSA4 quando ITUB4 são ações do Itaú Unibanco. Apesar de existirem algumas diferenças teóricas entre ambos, na prática são muito correlacionados. Portanto, neste caso, optamos por manter apenas o ITUB4.

Removemos também o papel 'BPAC11' (BTG Pactual), por conter um histórico muito pequeno (a partir de 2017). Ficamos, então, com 4 ações (tickers) para fazermos nossas análises iniciais.

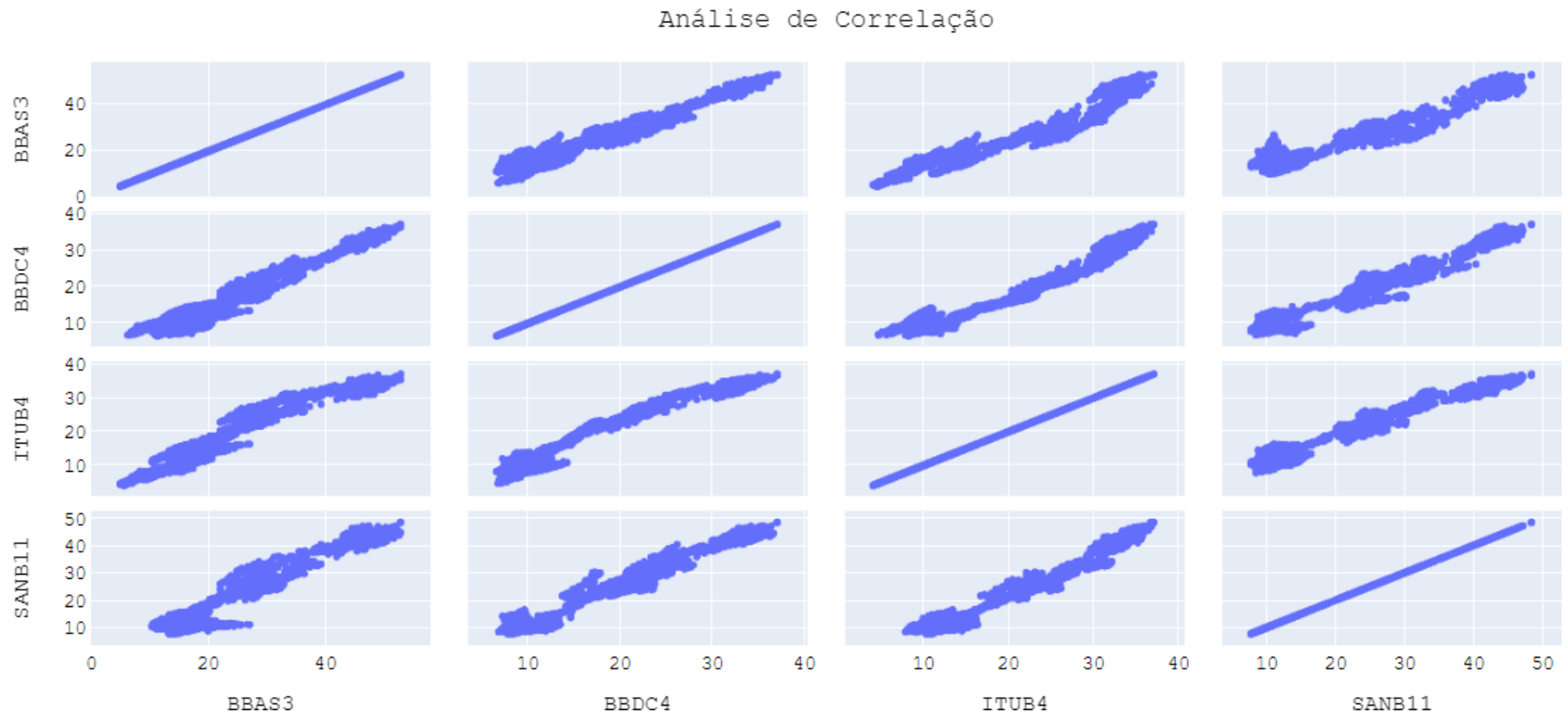
4. Análise Exploratória

4. Análise Exploratória



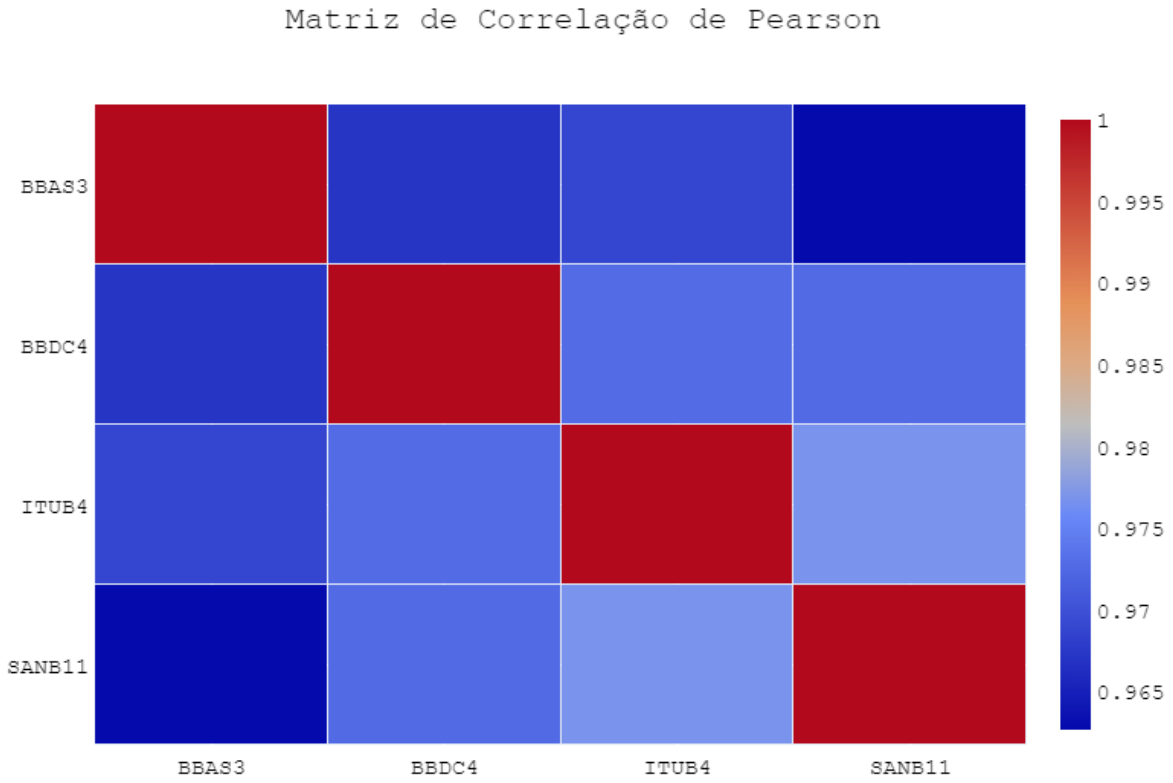
Observamos o histórico de preços das 4 ações (Banco do Brasil, Bradesco, Itaú Unibanco e Santander). Os papéis de Banco do Brasil e Itaú estão presentes desde o começo desta série temporal, isto é, 2005. A ação do Bradesco teve início em 2008 e a do Santander próximo de 2010. Neste gráfico, dois pontos chamam atenção: 1. As séries parecem estar correlacionadas (vamos confirmar isso mais pra frente); e 2. O papel do Banco do Brasil parece ter uma variância maior ao longo do tempo.

4. Análise Exploratória



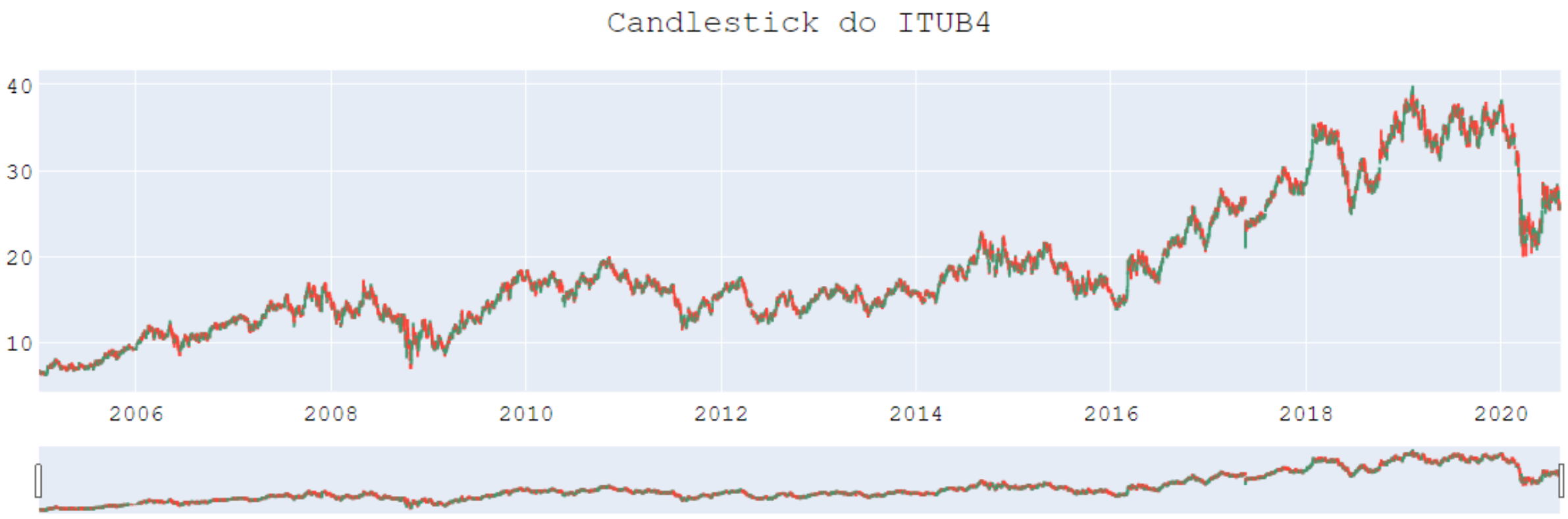
Estes gráficos nos mostram que de fato as cotações entre as instituições financeiras estão fortemente correlacionadas (positivamente). Nota-se, também, que os bancos privados (Itaú Unibanco, Bradesco e Santander) estão mais correlacionados entre si do que com o banco público (Banco do Brasil.

4. Análise Exploratória



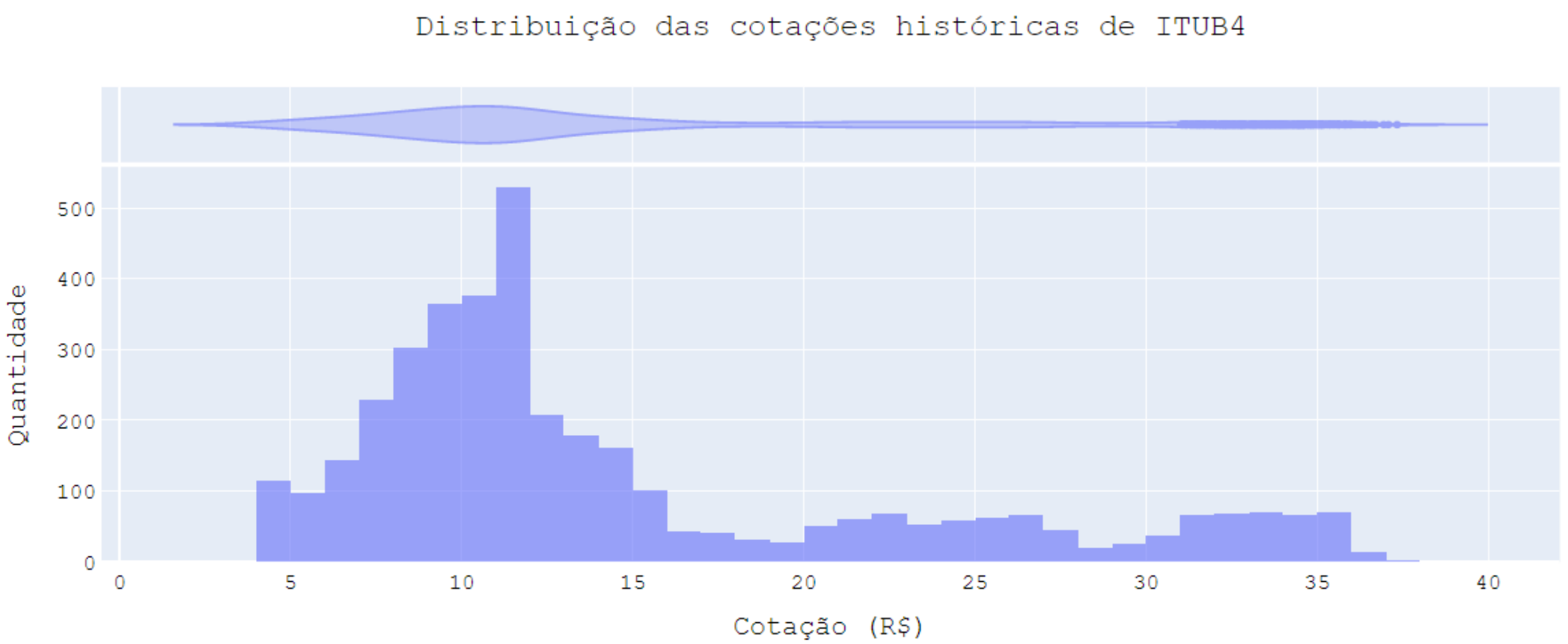
Este mapa de calor reforça os dois pontos que observamos no gráfico anterior. Nota-se que as correlações entre a ação do Banco do Brasil (BBAS3) com os demais papéis estão representadas com um azul mais escuro que, segundo a barra de cores, significa uma correlação menor, apesar de ainda ser uma correlação fortíssima.

4. Análise Exploratória



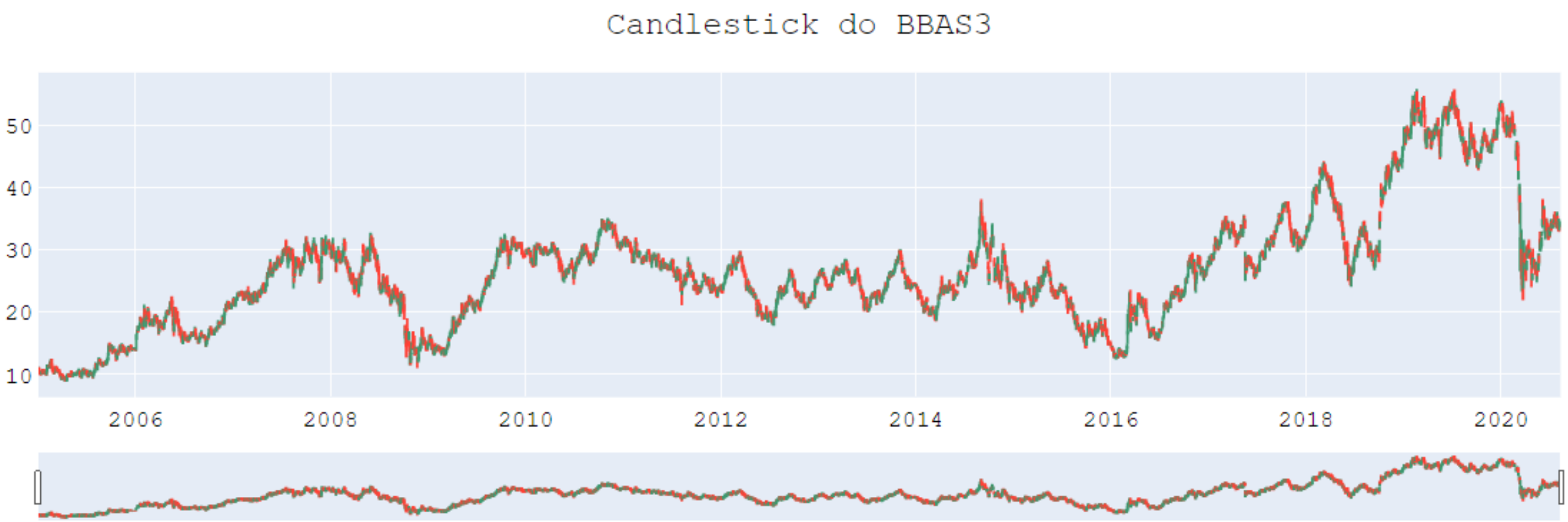
Aqui, vemos o gráfico do tipo Candlestick do papel ITUB4. Cada dia é representado por um boxplot, que informa 4 situações de preço da ação: 1. Preço de abertura; 2. Preço máximo; 3. Preço mínimo; e 4. Preço de fechamento. Os pontos em vermelho representam dias em que o preço de abertura foi maior do que o preço de fechamento, ou seja, indica uma desvalorização da ação. Já os pontos verdes, representam o contrário, isto é, dias em que o preço da ação subiu.'

4. Análise Exploratória



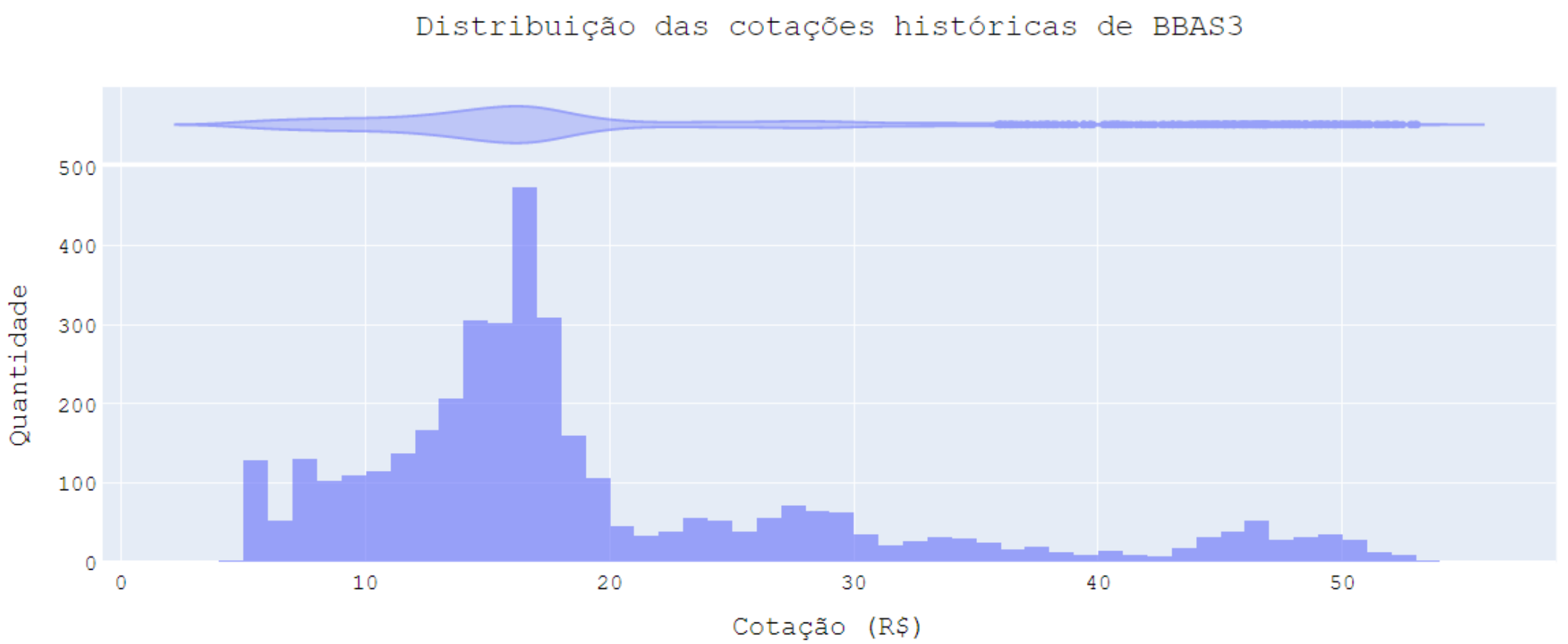
Acima vemos a distribuição das cotações históricas da empresa ITUB4. Nota-se pelo histograma que, ao longo do tempo, os preços ficaram concentrados entre R$ 5 e R 15. Comparando com o candlestick visto no slide anterior, é possível concluir que essa concentração deve-se ao fato de que a ação ficou muito tempo (de 2005 até 2014) com cotações nesta faixa.

4. Análise Exploratória



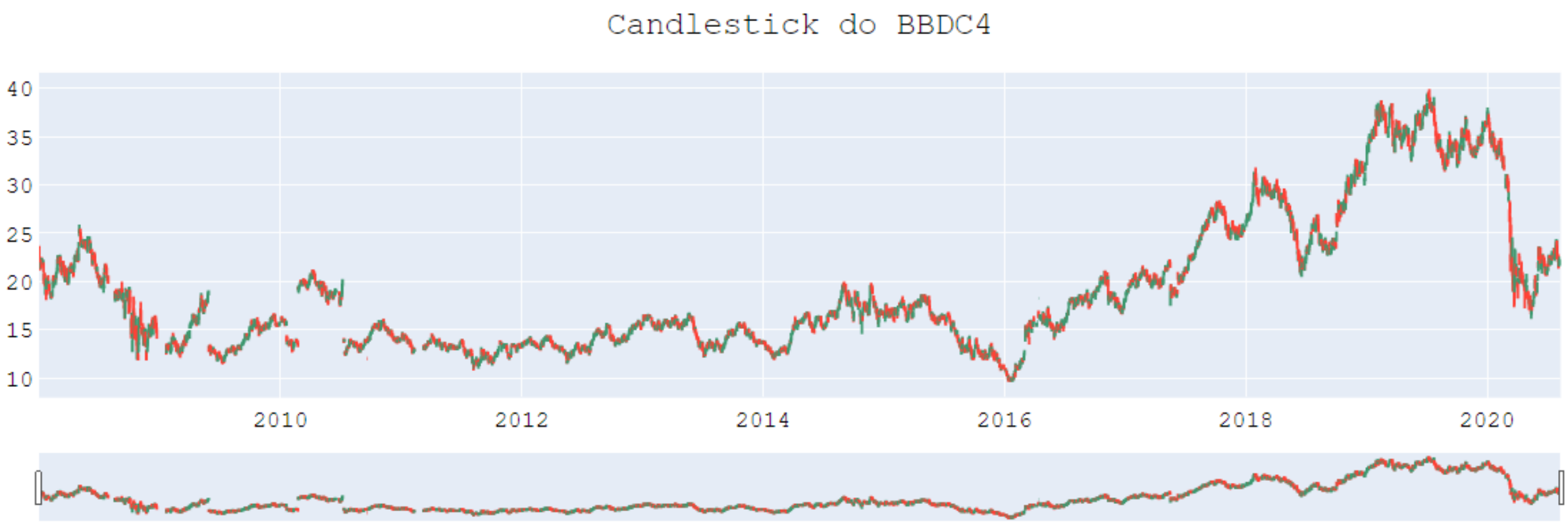
Aqui, vemos o gráfico do tipo Candlestick do papel BBAS3. Cada dia é representado por um boxplot, que informa 4 situações de preço da ação: 1. Preço de abertura; 2. Preço máximo; 3. Preço mínimo; e 4. Preço de fechamento. Os pontos em vermelho representam dias em que o preço de abertura foi maior do que o preço de fechamento, ou seja, indica uma desvalorização da ação. Já os pontos verdes, representam o contrário, isto é, dias em que o preço da ação subiu.

4. Análise Exploratória



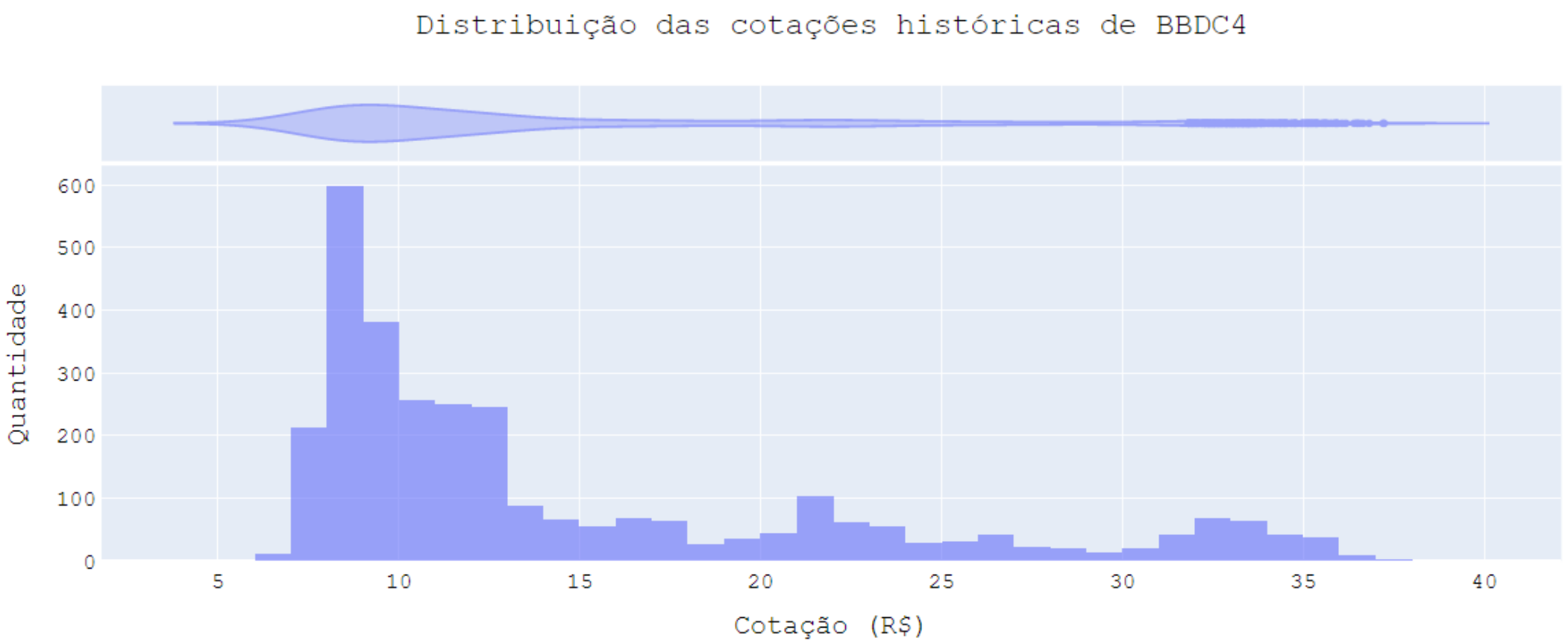
Acima vemos a distribuição das cotações históricas da empresa BBAS3. Nota-se pelo histograma que, ao longo do tempo, os preços ficaram concentrados entre R$ 5 e R 20. Comparando com o candlestick visto no slide anterior, é possível concluir que essa concentração deve-se ao fato de que a ação ficou muito tempo (de 2005 até 2016) com cotações nesta faixa. Comparando com a empresa anterior (ITUB4), nota-se que a distribuição é mais larga, o que indica maior variância e, portanto, maior oscilação de preço ao longo do tempo.

4. Análise Exploratória



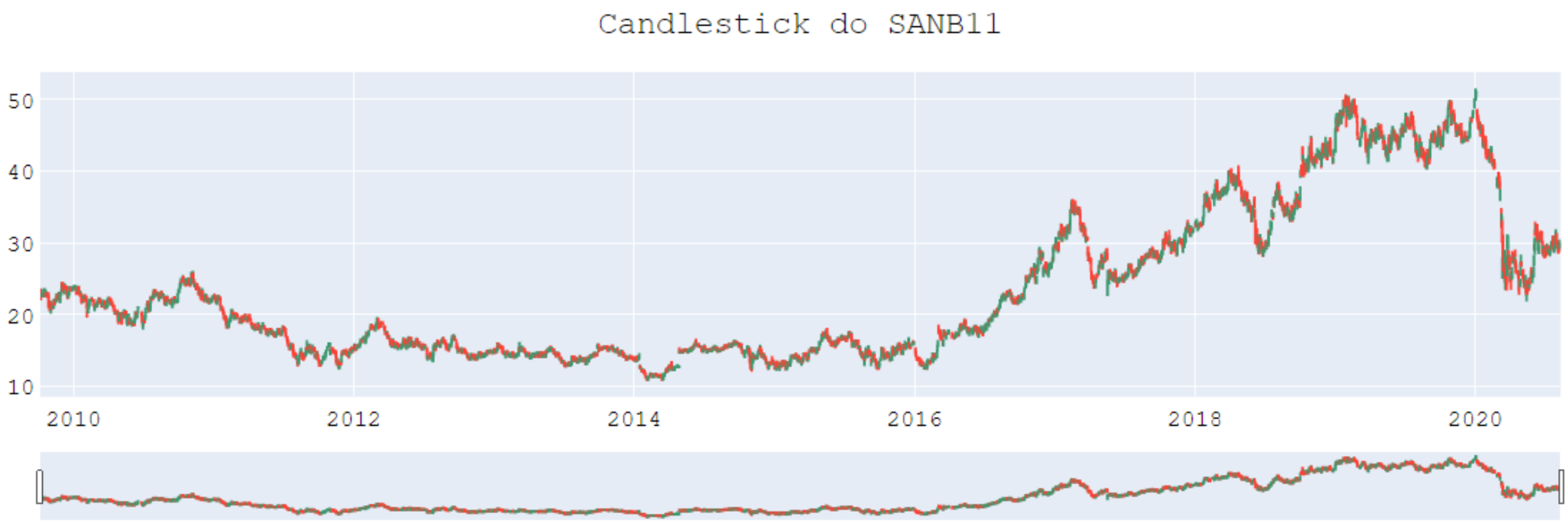
Aqui, vemos o gráfico do tipo Candlestick do papel BBDC4. Cada dia é representado por um boxplot, que informa 4 situações de preço da ação: 1. Preço de abertura; 2. Preço máximo; 3. Preço mínimo; e 4. Preço de fechamento. Os pontos em vermelho representam dias em que o preço de abertura foi maior do que o preço de fechamento, ou seja, indica uma desvalorização da ação. Já os pontos verdes, representam o contrário, isto é, dias em que o preço da ação subiu.

4. Análise Exploratória



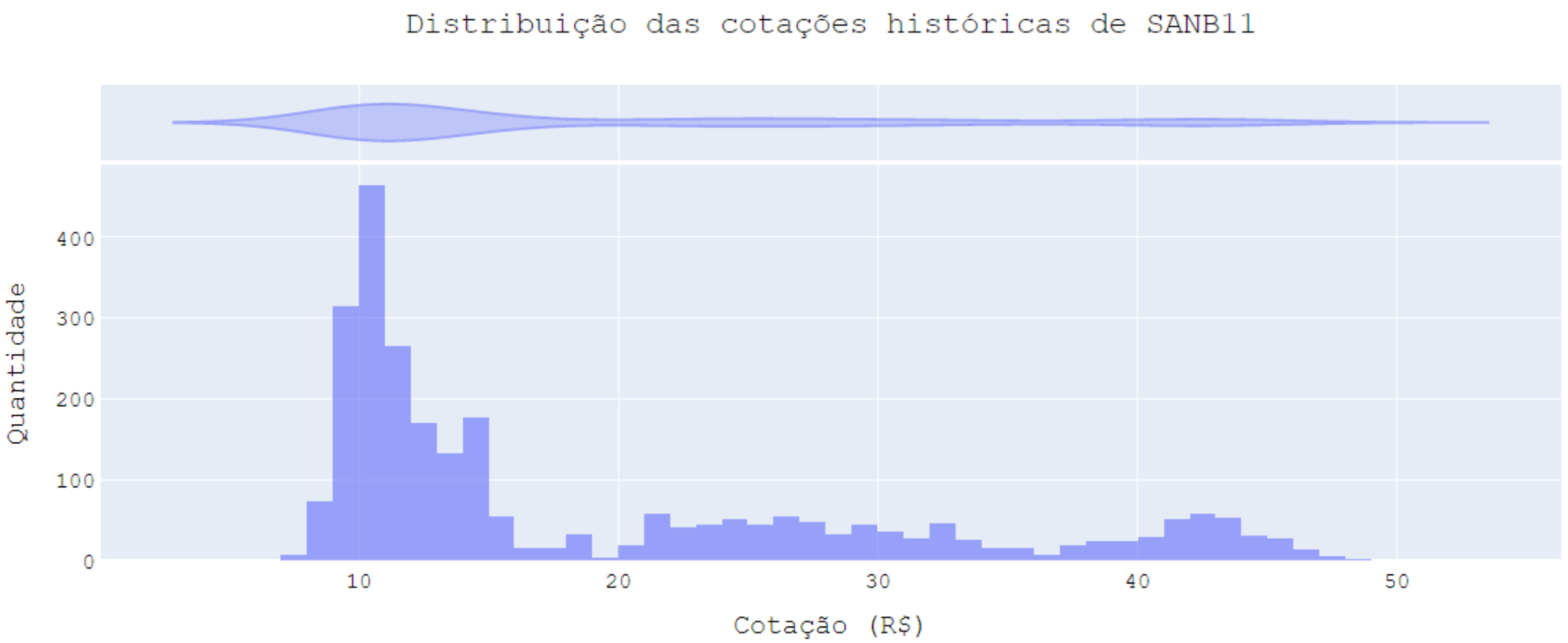
Acima vemos a distribuição das cotações históricas da empresa BBAS3. Nota-se pelo histograma que, ao longo do tempo, os preços ficaram concentrados entre R$ 7 e R 15. Comparando com o candlestick visto no slide anterior, é possível concluir que essa concentração deve-se ao fato de que a ação ficou muito tempo (de 2009 até 2017) com cotações nesta faixa.

4. Análise Exploratória



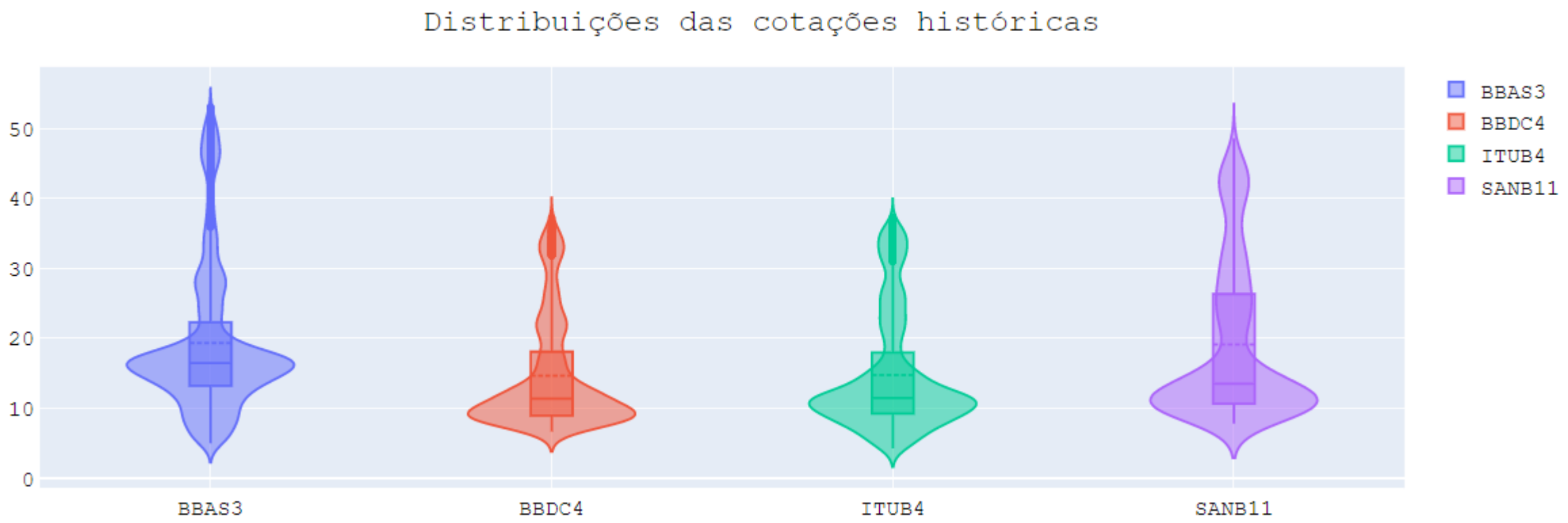
Aqui, vemos o gráfico do tipo Candlestick do papel SANB11. Cada dia é representado por um boxplot, que informa 4 situações de preço da ação: 1. Preço de abertura; 2. Preço máximo; 3. Preço mínimo; e 4. Preço de fechamento. Os pontos em vermelho representam dias em que o preço de abertura foi maior do que o preço de fechamento, ou seja, indica uma desvalorização da ação. Já os pontos verdes, representam o contrário, isto é, dias em que o preço da ação subiu.

4. Análise Exploratória



Acima vemos a distribuição das cotações históricas da empresa BBAS3. Nota-se pelo histograma que, ao longo do tempo, os preços ficaram concentrados entre R$ 9 e R 15. Comparando com o candlestick visto no slide anterior, é possível concluir que essa concentração deve-se ao fato de que a ação ficou muito tempo (de 2010 até 2016) com cotações nesta faixa.

4. Análise Exploratória

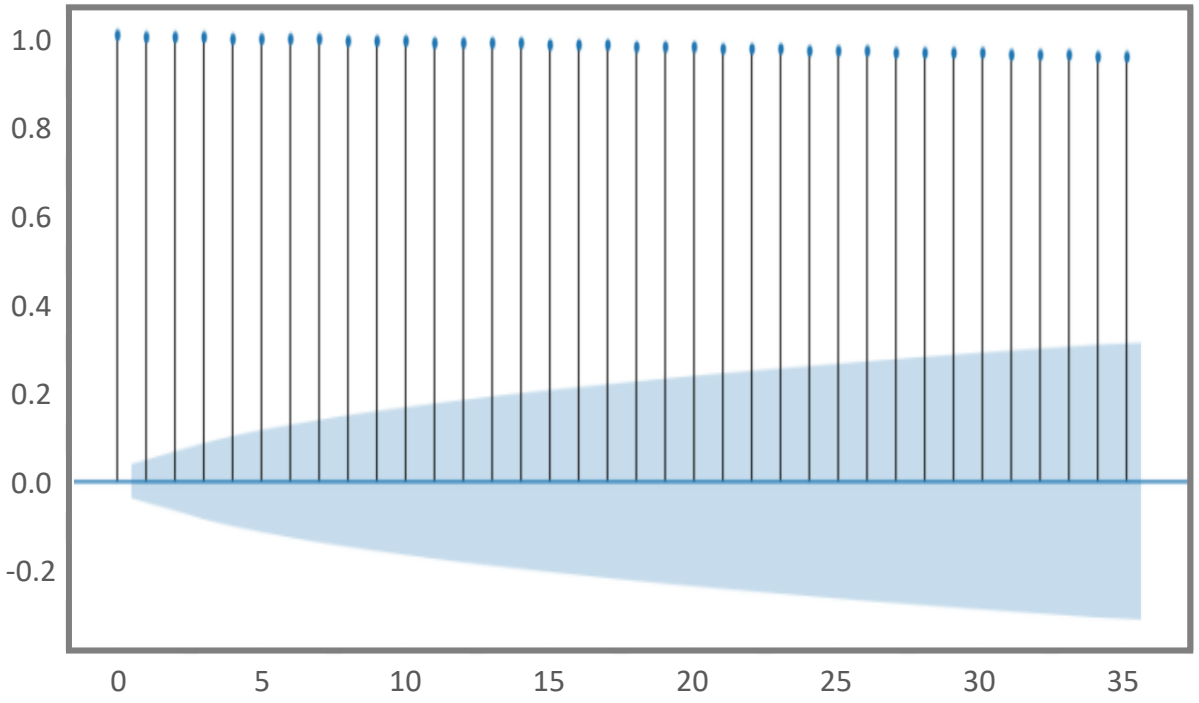


Neste gráfico podemos comparar as distribuições das 4 ações juntas. Observa-se que os bancos Itaú (ITUB4) e Bradesco (BBDC4) possuem distribuições bem semelhantes e com menor variância em relação às outras instituições. O BBAS3 possui uma variância maior, provavelmente em função de ser um banco público e estar mais sujeito aos impactos de notícias relacionadas à política. Já no papel do Santander, também notamos uma alta variância, porém, nesse caso, é mais provável que seja por conta do momento em que esta empresa abriu seu capital (2009).

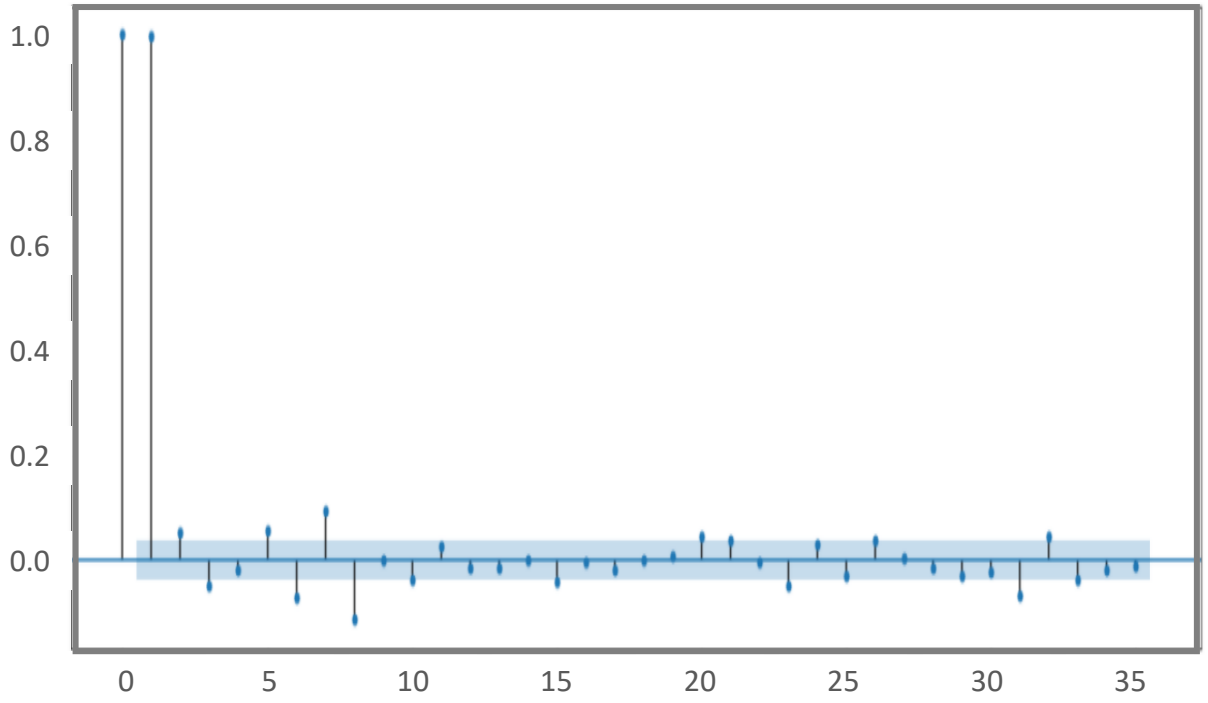
4. Análise Exploratória

ITUB4

Autocorrelação



Autocorrelação Parcial

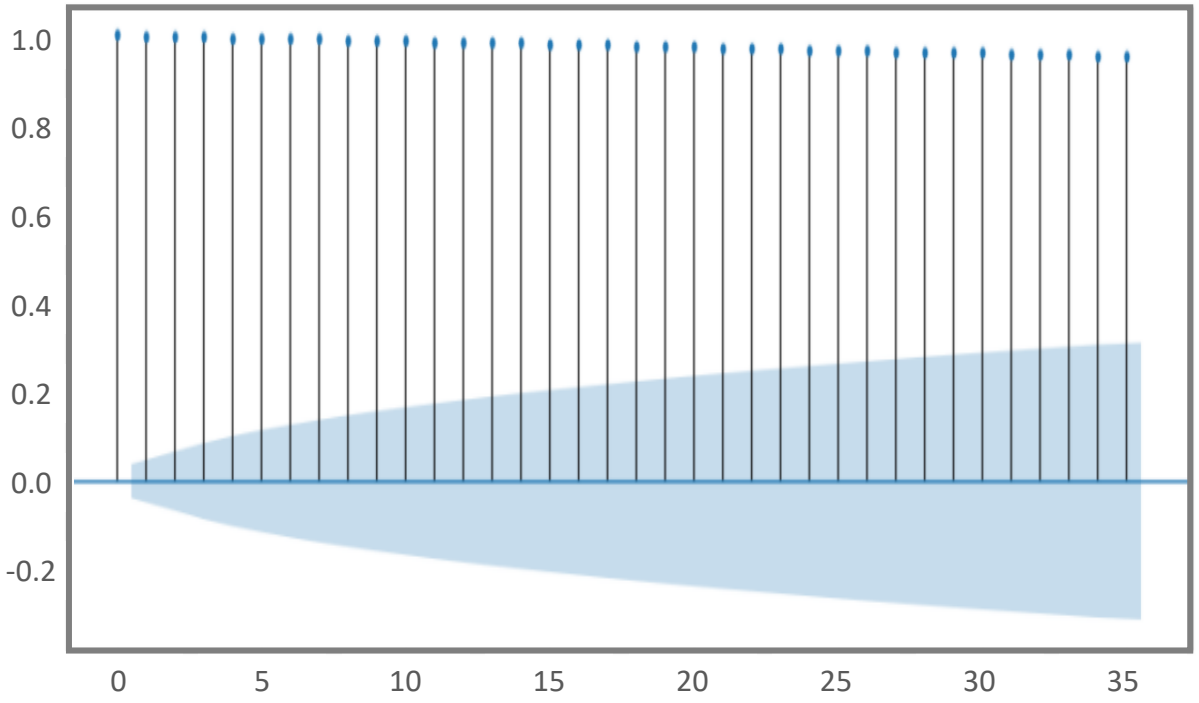


Nesta quadro, podemos analisar a autocorrelação da ação ITUB4 (Itaú Unibanco). A autocorrelação é a correlação da variável temporal com ela mesma em períodos anteriores. Por exemplo: qual a influência que um preço de 2 dias atrás (lag 2) possui no preço de hoje? A autocorrelação parcial também analisa a autocorrelação, porém descontando as influências dos períodos seguintes. Por exemplo: qual a influência que um preço de 2 dias atrás (lag 2) possui no preço de hoje, descontada a influência do lag 1?

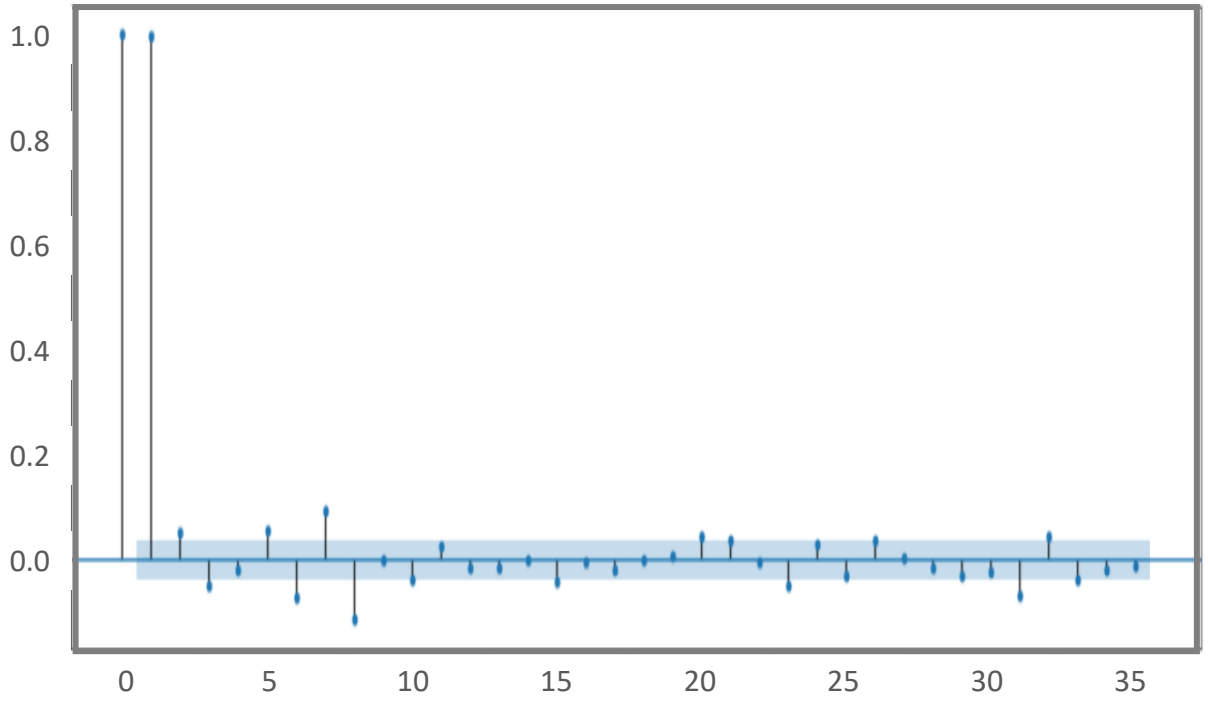
4. Análise Exploratória

ITUB4

Autocorrelação



Autocorrelação Parcial



Essas informações são relevantes para definirmos a ordem do nosso modelo ARIMA, que utilizaremos na parte de modelagem tradicional. Notamos que de fato os dias anteriores possuem forte influência nos preços das ações do dia 0. Contudo, estamos analisando uma série não-estacionária. Precisamos transformá-la em estacionária e verificar se a autocorrelação se mantém. Como as séries temporais de todos os papéis são fortemente correlacionadas, não plotaremos os gráficos de autocorrelação dos demais, pois serão muito parecidos.

5. Tratamento da Base

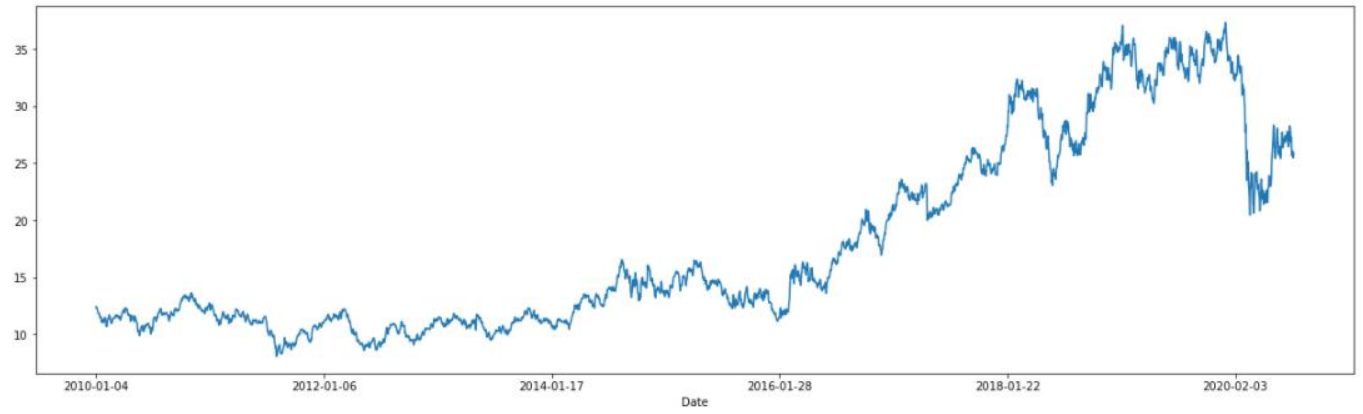
5. Tratamento da Base

Transformando a série em estacionária

Não Estacionária

Date	ITUB4	
04/08/2020	R$	25,69
05/08/2020	R$	25,67
06/08/2020	R$	26,03
07/08/2020	R$	25,48
10/08/2020	R$	25,81

ITUB4



Nossa série temporal é não estacionária, isto é, não possuem média e variância constante ao longo do tempo, além de apresentar uma tendência de crescimento. Logo, é necessário transformá-la em estacionária. Para isso, tiramos o que chamamos de "nível" da série. Isto é, ao invés de utilizarmos os preços em si, analisaremos as diferenças entre os preços. Essa diferença pode ser absoluta ou em percentual. Escolhemos trabalhar com as diferenças percentuais.

5. Tratamento da Base

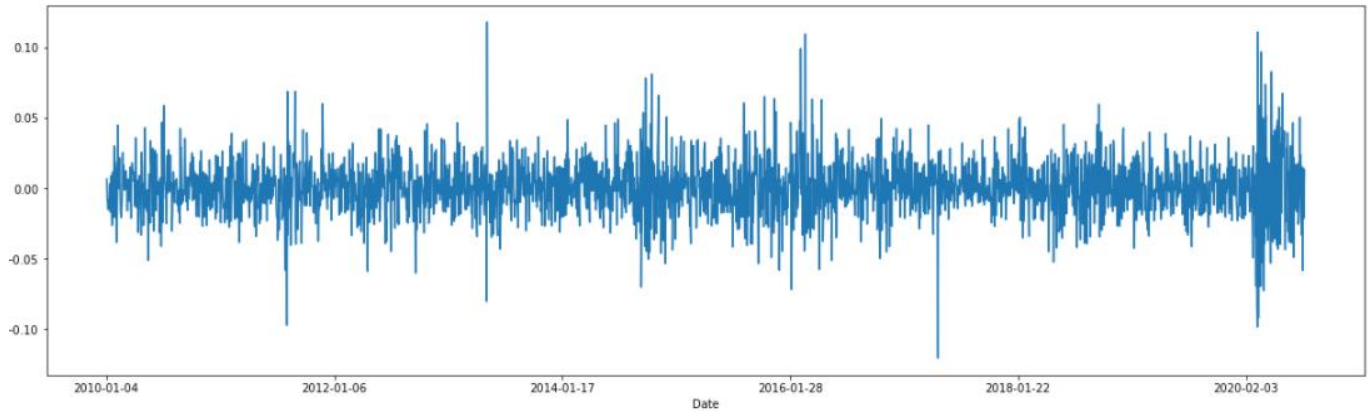
Transformando a série em estacionária

Estacionária

Date	ITUB4		ITUB4 (%)
04/08/2020	R$	25,69	-5,83%
05/08/2020	R$	25,67	-0,08%
06/08/2020	R$	26,03	1,40%
07/08/2020	R$	25,48	-2,11%
10/08/2020	R$	25,81	1,30%

(Lag 1)

ITUB4



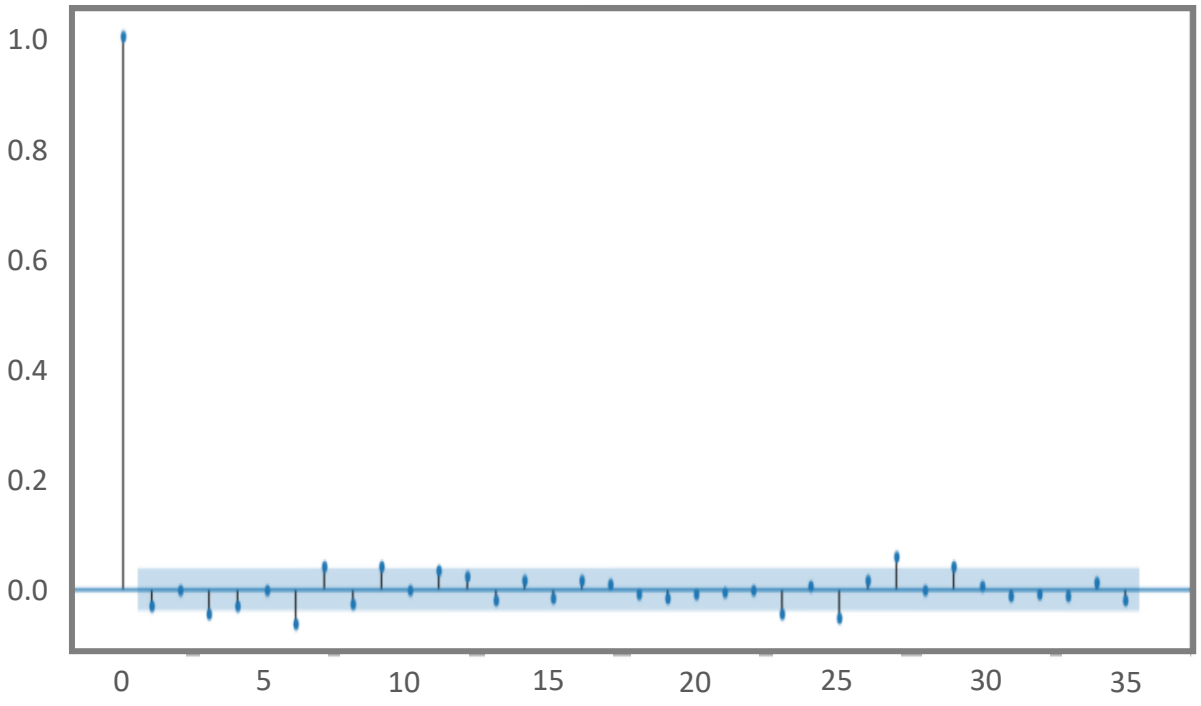
Teste de Dickey-Fuller → P-Value = 0.0

Nota-se que ao utilizar a diferença percentual em relação ao dia anterior (diferenciação de ordem 1), conseguimos transformar nossa série em estacionária. Isso pode ser confirmado pelo teste de Dickey-Fuller, que testa a hipótese da série não ser estacionária: como o p-value ficou abaixo do nível de significância de 5%, podemos rejeitar a hipótese de que a série é não-estacionária e, portanto, afirmar que ela é.

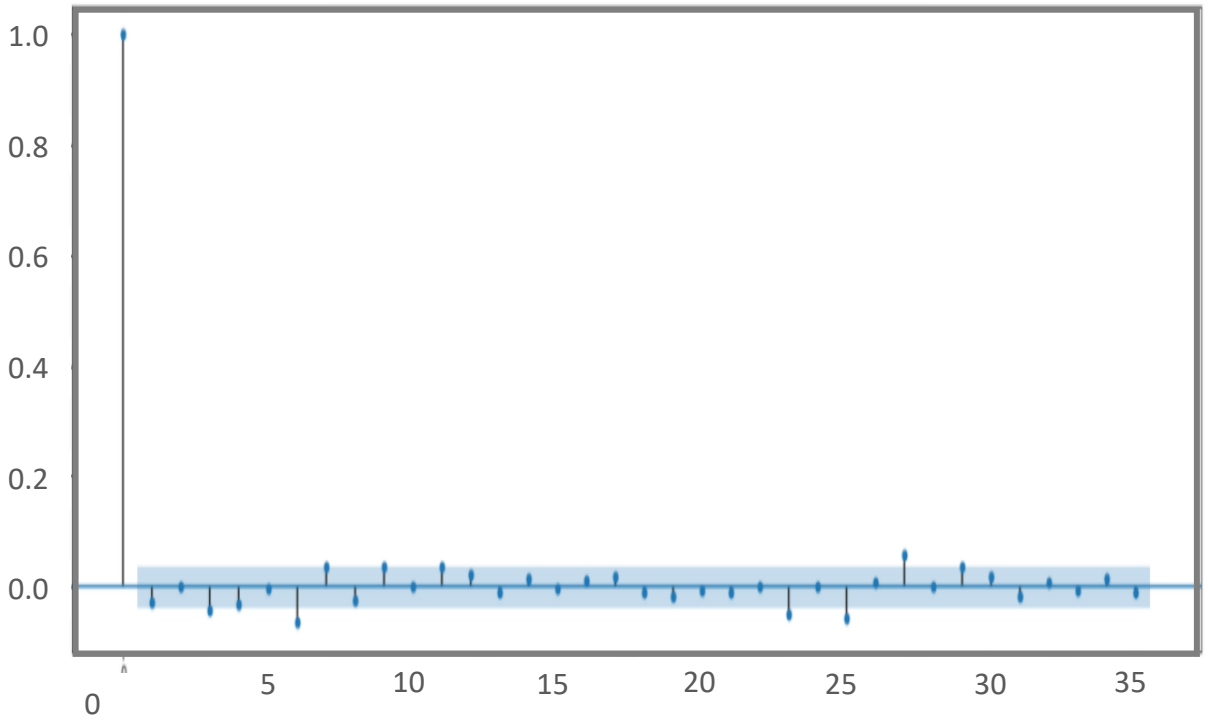
5. Tratamento da Base

ITUB4

Autocorrelação



Autocorrelação Parcial



Notamos que a autocorrelação na série estacionária é praticamente igual a zero, o que indica que a série (pelo menos na visão diária) é imprevisível, o que classificamos como White Noise (ou Ruído Branco). Por isso, é necessário que na modelagem testemos com diferentes agrupamentos temporais (diário, semanal, mensal, etc.), pois pode ser que desta forma consigamos encontrar algum padrão relevante e previsível.

5. Tratamento da Base

Definição do Período

Ticker	Data Início	Data Fim
BBAS3	03/01/2005	10/08/2020
BBDC4	02/01/2008	10/08/2020
ITUB4	03/01/2005	10/08/2020
SANB11	07/10/2009	10/08/2020

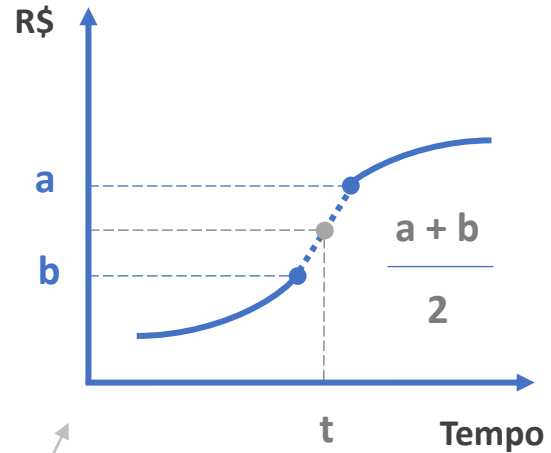
Período definido: a partir de 2010

As ações BBAS3 e ITUB4 possuem valores desde 2005. BBDC4, a partir de 2008. E SANB11, a partir do final de 2009. Portanto, vamos considerar de 2010 para frente. Assim, teremos um bom período para analisar e não consideramos dados muitos antigos, que tendem a ter pouca influência sobre os dados mais recentes. Desta forma, saímos de um shape de (3878, 4) para (2629, 4).

5. Tratamento da Base

Tratamento de Missings

Ticker	Qtd. Nulos	Perc. Nulos
BBAS3	13	0,49%
BBDC4	33	1,26%
ITUB4	12	0,45%
SANB11	11	0,42%



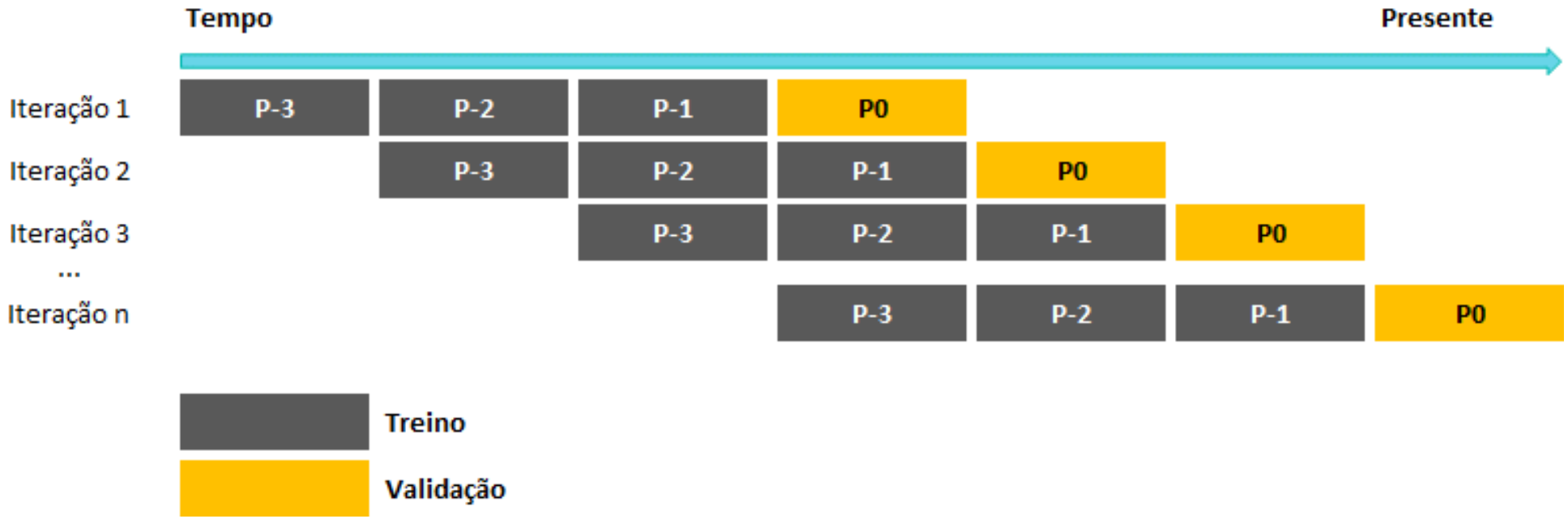
Método de tratamento: Interpolação Linear

Há algumas maneiras de se fazer tratamento de missings em séries temporais. Pode-se, por exemplo, utilizar o valor do momento anterior ou posterior. Porém, na maioria dos casos, esses métodos não são recomendados, principalmente quando há uma tendência clara na série temporal. Nos gráficos das ações, nota-se um comportamento tendencioso ao longo do tempo e, por isso, o método mais indicado para tratamento de missings é a interpolação linear.

6. Modelagem Tradicional

6. Modelagem Tradicional

Método de Validação – Janela Deslizante



Para validar os nossos modelos, escolhemos utilizar o método de "Janela Temporal Deslizante", que consiste em, para todas as iterações, utilizar os dados dos X períodos anteriores para treinar e predizer no momento seguinte. É possível também fazer testes utilizando o método "Janela de Expansão", porém não aplicaremos aqui.

6. Modelagem Tradicional

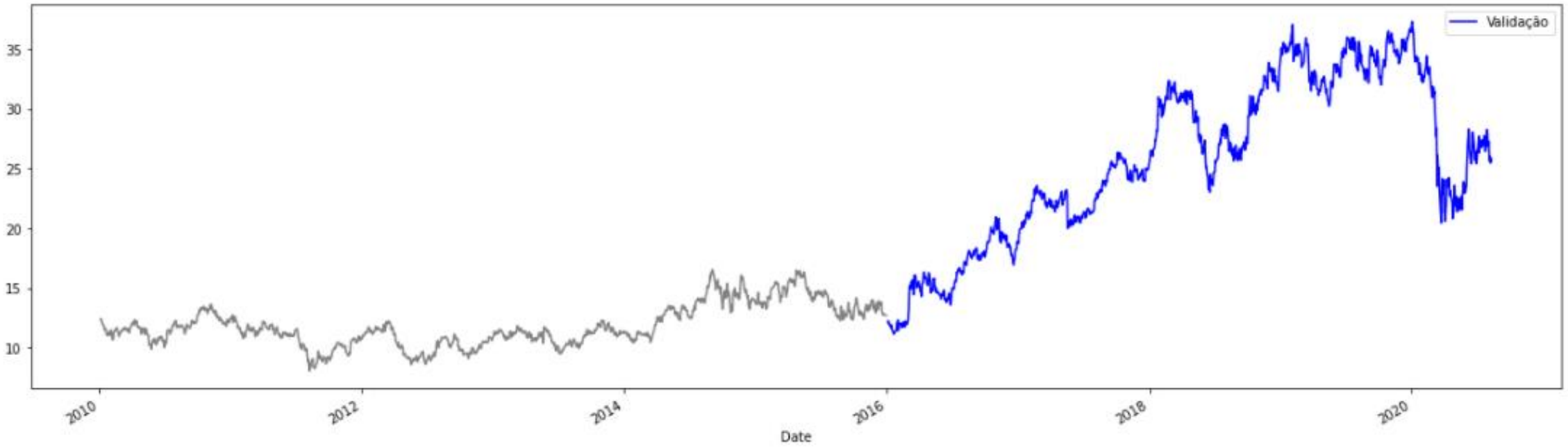
Separação em treino e validação

Separamos os dados em treino e validação conforme a estratégia acima. Neste exemplo, os dados estão agrupados por ano, porém esse agrupamento será feito com granularidade diária, semanal e mensal. Não optamos por agrupamentos mais "graúdos" por conta da pouca quantidade de dados que teríamos.

6. Modelagem Tradicional

Separação em treino e validação

ITUB4



Separamos os dados em treino e validação conforme a estratégia acima. Neste exemplo, os dados estão agrupados por ano, porém esse agrupamento será feito com granularidade diária, semanal e mensal. Não optamos por agrupamentos mais "graúdos" por conta da pouca quantidade de dados que teríamos.

6. Modelagem Tradicional

Resultados da Modelagem

ITUB4

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0234	0,0234	0,0234
ARIMA	0,0213	0,0446	0,0890
Linear Regression	0,0214	0,0446	0,0928
Random Forest	0,0217	0,0446	0,0903

BBDC4

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0292	0,0292	0,0292
ARIMA	0,0247	0,0494	0,1051
Linear Regression	0,0246	0,0505	0,1061
Random Forest	0,0249	0,0585	0,1055

BBAS3

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0308	0,0308	0,0308
ARIMA	0,0303	0,0677	0,1401
Linear Regression	0,0305	0,0689	0,1448
Random Forest	0,0326	0,0698	0,1499

SANB11

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0263	0,0263	0,0263
ARIMA	0,0240	0,0515	0,1037
Linear Regression	0,0239	0,0520	0,1141
Random Forest	0,0237	0,0540	0,1134

Nota-se que o modelo ARIMA obteve melhores resultados. Também é possível perceber que o agrupamento diário apresenta erros menores. Nesta modelagem tradicional, não fizemos tunning dos hiperparâmetros dos modelos, o que poderia ter melhorado os resultado. É bom ressaltar também que aqui estamos medindo apenas o erro, e não retorno financeiro. Nem sempre o modelo que apresenta o menor erro nas previsões será o mais lucrativo. É necessário fazer essa medição, também, e comparar as abordagens.

7. Modelagem Avançada

7. Modelagem Avançada

Resultados da Modelagem

ITUB4

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0234	0,0234	0,0234
ARIMA	0,0213	0,0446	0,0890
Linear Regression	0,0214	0,0446	0,0928
Random Forest	0,0217	0,0446	0,0903
Light GBM	0,0213	0,0444	0,0884
LSTM	0,0227	0,0450	0,0893
Prophet	0,0219	0,0463	0,1068

BBDC4

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0292	0,0292	0,0292
ARIMA	0,0247	0,0494	0,1051
Linear Regression	0,0246	0,0505	0,1061
Random Forest	0,0249	0,0585	0,1055
Light GBM	0,0246	0,0495	0,1043
LSTM	0,0250	0,0500	0,1052
Prophet	0,0252	0,0511	0,1209

BBAS3

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0308	0,0308	0,0308
ARIMA	0,0303	0,0677	0,1401
Linear Regression	0,0305	0,0689	0,1448
Random Forest	0,0326	0,0698	0,1499
Light GBM	0,0303	0,0677	0,1404
LSTM	0,0309	0,0675	0,1393
Prophet	0,0308	0,0697	0,1703

SANB11

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0263	0,0263	0,0263
ARIMA	0,0240	0,0515	0,1037
Linear Regression	0,0239	0,0520	0,1141
Random Forest	0,0237	0,0540	0,1134
Light GBM	0,0240	0,0515	0,1031
LSTM	0,0404	0,0562	0,1029
Prophet	0,0247	0,0531	0,1169

Percebe-se que os modelos mais sofisticados não superaram as abordagens mais simples. Dos 3 modelos adicionados nesta etapa, apenas o Light GBM chegou mais próximo do desempenho dos modelos mais tradicionais. Já o LSTM e o Prophet tiveram performance bem inferior.

7. Modelagem Avançada

Resultados da Modelagem

Média

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	0,0274	0,0274	0,0274
ARIMA	0,0251	0,0533	0,1095
Linear Regression	0,0251	0,0540	0,1145
Random Forest	0,0257	0,0567	0,1148
Light GBM	0,0251	0,0533	0,1091
LSTM	0,0297	0,0547	0,1092
Prophet	0,0256	0,0551	0,1287

Ranking

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	6	1	1
ARIMA	2	2	4
Linear Regression	3	4	5
Random Forest	5	7	6
Light GBM	1	3	2
LSTM	7	5	3
Prophet	4	6	7

Neste quadro podemos ver mais claramente os desempenhos dos modelos. Do lado esquerdo, calculou-se a média dos erros das 4 séries temporais para cada modelo e agrupamento. Nesta tabela, notamos que as cores mais avermelhadas ficaram concentradas nos modelos LSTM, Prophet e Random Forest. Do lado direito, vemos o ranking das performances médias. O baseline, apesar de ter tido a primeira colocação em dois agrupamentos, ficou em sexto em um deles. O Light GBM se mostrou mais estável do ponto de vista "posição".

Resultados da Modelagem

Dispersão

Modelo	RMSE_Diário	RMSE_Semanal	RMSE_Mensal
Baseline	9,4%	0,0%	0,0%
ARIMA	0,0%	94,3%	299,1%
Linear Regression	0,2%	96,9%	317,3%
Random Forest	2,6%	106,7%	318,4%
Light GBM	0,0%	94,3%	297,6%
LSTM	18,7%	99,3%	298,0%
Prophet	2,3%	100,8%	369,3%

Neste quadro, vemos os erros percentuais relativos ao primeiro colocado. Olhando apenas para o quadro de ranking, tendemos a concluir que o modelo Light GBM é mais estável. Porém, quando analisamos neste quadro o erro percentual relativo ao modelo de melhor performance, vemos que quando o Baseline ficou em 6º colocado, a diferença percentual do erro em relação ao primeiro colocado (Light GBM) foi de apenas 9,4%. Por outro lado, quando o Baseline ficou em 1º colocado, o Light GBM ficou com erros muito mais dispersos: 94,3% e 297,6%.

8. Conclusões

8. Conclusões

Concluímos que **os modelos mais sofisticados não se mostraram superiores em relação ao baseline e os modelos mais básicos**. Isso pode ter acontecido por basicamente 3 motivos:

1. **Os modelos mais sofisticados**, por serem mais flexíveis e se ajustarem melhor aos dados, também **são mais sensíveis aos ruídos dos dados, o que pode causar overfitting**;

2. **Os dados utilizados são insuficientes** para que os modelos identifiquem padrões e, portanto, consigam prever comportamentos futuros;

3. **Não há padrões nas séries temporais do mercado financeiro**, o que chamamos de "Random Walk" (ou "Passeio Aleatório")

Um ponto importante, já citado anteriormente, é que **o desempenho avaliado foi uma métrica de erro, e não o retorno financeiro de cada modelo**. **Nem sempre o modelo com o menor erro será aquele mais lucrativo.**

Como próximos passos, a fim de melhorar os resultados, sugere-se:

- Testar outros hyperparâmetros
- **Testar outros modelos (ex.: Redes Convolucionais 1D)**
- Testar outros agrupamentos
- **Utilizar técnicas de séries temporais multivariadas**
- **Capturar novos dados**

Séries Temporais aplicadas ao Mercado Financeiro

Modelo preditivo de preços das ações da B3

22/10/2020