UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

Thaii Céu

TRABALHO DE SÉRIES TEMPORAIS – ESTUDO DA EXPORTAÇÃO DO CAFÉ ARÁBICA

PRESIDENTE PRUDENTE
2023

SUMÁRIO

Introdução

O café Arábica é uma espécie originária da Etiópia, sendo uma das primeiras a serem cultivadas. Seu nome "Arábica" é atribuído à propagação de seu cultivo pelos árabes na região da Arábia há milhares de anos. Atualmente, essa espécie é cultivada globalmente. No Brasil, a produção de desse tipo de café se concentra principalmente em Minas Gerais, São Paulo, Paraná e Bahia. O Espírito Santo, embora produza menos dele, se destaca, junto com Rondônia, na produção de café robusta (ou Conillon).

Essa espécie é mais sensível e vulnerável a pragas e condições climáticas adversas. Os pés de café arábica prosperam melhor em altitudes acima de 800 metros, em climas amenos, que são considerados ideais para produzir grãos de alta qualidade. Por isso, os grãos têm um valor de mercado mais alto.

Ele é reconhecido por ser um grão especial, de cultivo mais complexo, com uma variedade de aromas intensos e sabores diversificados, além de diferentes níveis de corpo e acidez.

O café é um dos pilares da economia brasileira desde o século XIX. Além de ser um grande consumidor, o Brasil é o maior produtor e exportador de café no mundo. Em 2023, o país exportou 39,247 milhões de sacas de 60 kg de café, um volume praticamente estável em relação às 39,410 milhões de sacas exportadas em 2022, apresentando uma leve queda de 0,4%. Em termos de receita cambial, houve uma redução de 13% em comparação ao ano anterior, com as exportações gerando US$ 8,041 bilhões. Esses dados são do relatório estatístico mensal do Conselho dos Exportadores de Café do Brasil (Cecafé).

No ano passado, o café arábica foi o mais exportado, com 30,818 milhões de sacas, representando 78,5% do total. A variedade canéfora (conilon + robusta) teve 4,708 milhões de sacas exportadas, correspondendo a 12% do total. Além disso, o segmento de café solúvel exportou 3,671 milhões de sacas (9,4%), e o café torrado e moído teve 50.377 sacas exportadas (0,1%).

Sendo assim, escolhi analisar a exportação do volume do café Arábica (entre janeiro de 2019 a dezembro de 2023), criando modelos e escolhendo o melhor para esse tipo de série.

Base de Dados

A base de dados foi obtida do relatório completo de exportação de café do site CECAFE (Conselho dos Exportadores de Café do Brasil). Eles disponibilizam mensalmente um relatório, mas podemos baixar o completo, onde nele existe todos os registros de exportação desde 1991 até o ano atual. Na base, além disso, existe 3 variáveis (Volume exportado por mês/ano, Receita Cambial por mês/ano, e Preço médio por mês/ano), e dentro de cada uma, existe os dados dos 4 tipos de café (Arábica, Conillon, Solúvel e Torrado)

Como existe registros antigos de exportação, eu retirei da base principal somente os anos que usaria (janeiro de 2019 a dezembro de 2023) e a variável de interesse (Volume exportado por mês/ano). Nesta base que retirei, existe o volume exportado os dados dos 4 tipos de café, porém, no trabalho apresentado, irei utilizar apenas uma série, que será a do Arábica.

Tabela 1: Base de dados do Volume exportado dos quatro tipos de café

> head(Volume)

A tibble: 6 × 5

Mês/Ano	**Conillon**	**Arábica**	**Torrado**	**Solúvel**
2019-01-31	150609	3073902	1477	244684
2019-02-28	212003	3051113	1220	310465
2019-03-31	190880	2557882	2202	366311
2019-04-30	240371	2706379	2319	318296
2019-05-31	462472	3025754	1634	396854
2019-06-30	383915	2355971	4099	358811

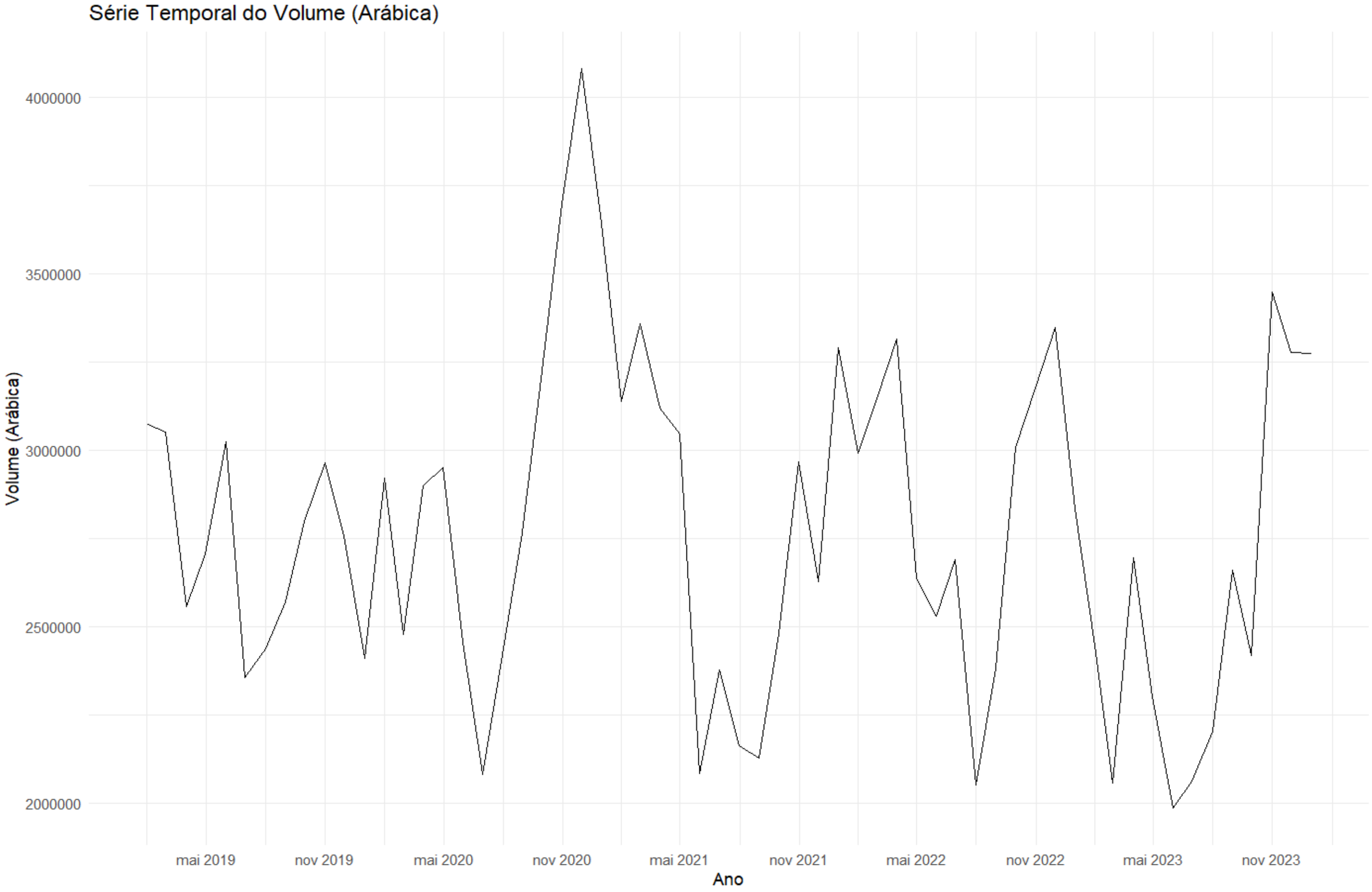
Existe, na base total, 60 observações de 5 variáveis. Acima, só estou mostrando as 6 primeiras observações, para ter uma visão de como está a base de dados. Só nos primeiros 6 dados, podemos ver que o arábica é o mais exportado. Para toda a análise, utilizei o programa R.

ANÁLISE DA SÉRIE

Estacionariedade

Antes de iniciarmos a análise, precisamos saber se a série que estamos estudando é estacionária ou não. Para isso, vamos observar o gráfico da série (de janeiro de 2019 a dezembro de 2023):

Figura 1: Gráfico da série de exportação do Volume do café Arábica



Observando apenas a série, podemos criar a hipótese de que ela é estacionária, já que ela parece ser constate a todo tempo. Para termos certeza, iremos utilizar o teste de Dickey-Fuller Aumentado, que verifica se a série é estacionária ou não. As hipóteses desse teste são:

H0: A série não é estacionária (Existe raiz unitária)

Ha: A série é estacionária (Não existe raiz unitária)

Se o valor-p do teste for menor que o nível de significância, rejeitamos a hipótese nula em favor da hipótese alternativa.

Logo:

> adf.test(Volume$Arábica)

Augmented Dickey-Fuller Test

data: Volume$Arábica

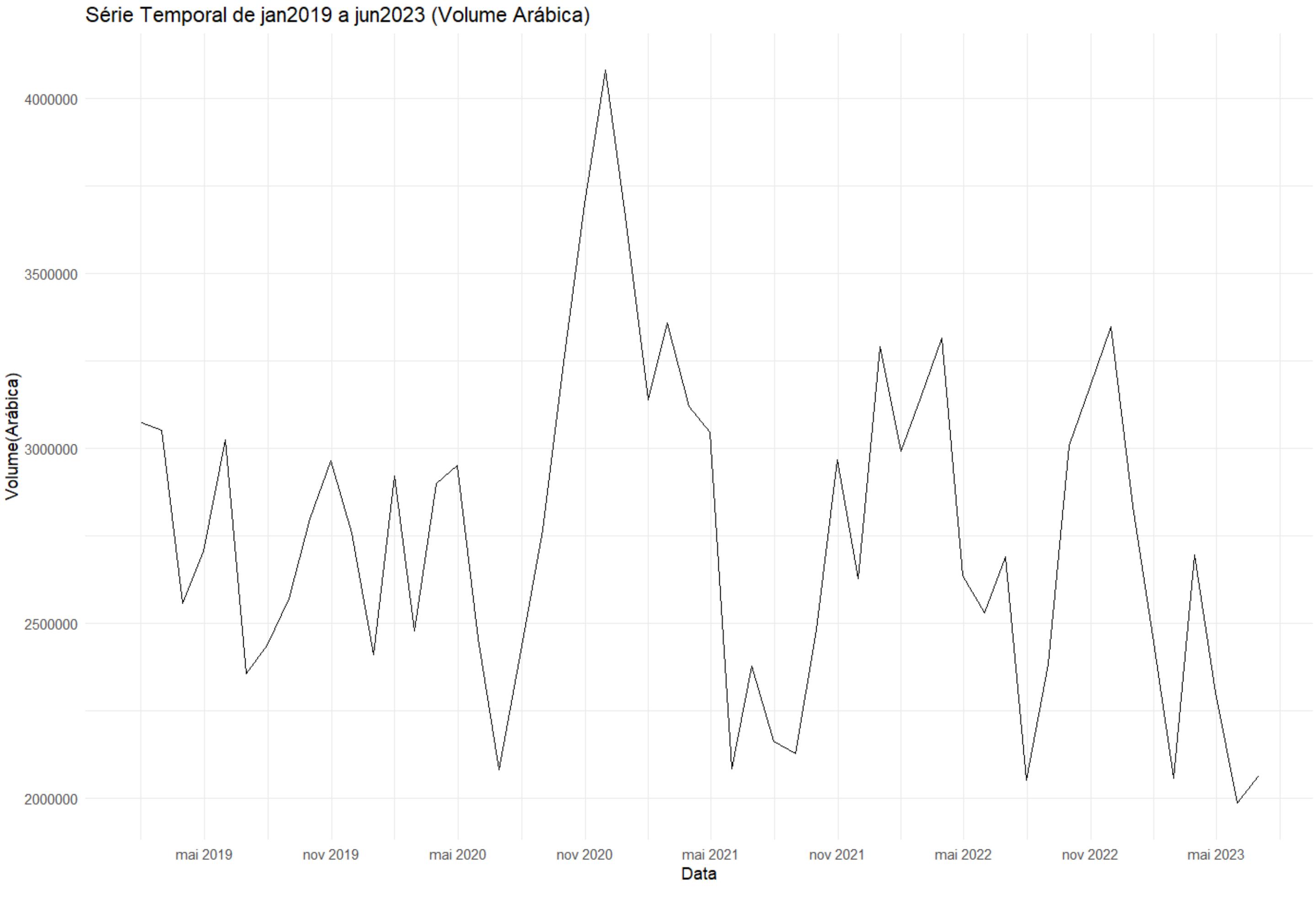
Dickey-Fuller = -3.9057, Lag order = 3, p-value = 0.01982

alternative hypothesis: stationary

Aderindo nosso nível de significância 0.05, e sendo nosso p-valor menor que ele, podemos rejeitar H0 em favor da Ha e dizer que ela é estacionária!

Como queremos criar um modelo de previsão para a série, vamos tirar as últimas 6 observações, ficando com uma série de janeiro de 2019 a junho de 2023. Logo:

Figura 2: Gráfico da série de exportação do Volume do café Arábica



Observando essa série, podemos criar a hipótese de que ela é estacionária. Vamos confirmar no teste:

```
> adf.test(Vol$Arábica)

Augmented Dickey-Fuller Test

data:  Vol$Arábica
Dickey-Fuller = -3.4618, Lag order = 3, p-value = 0.05552
alternative hypothesis: stationary
```

Por pouco, pelo teste, p-valor parece ser maior que nosso alfa = 0,05. Porém, podemos usar uma outra alternativa para vermos se é estacionária ou não:

```
> model1 = lm(Vol$Arábica ~ Vol$`Mês/Ano`)
> summary(model1) # Deu estácionaria!!

Call:
lm(formula = Vol$Arábica ~ Vol$`Mês/Ano`)

Residuals:
    Min      1Q  Median      3Q     Max
-721446 -349678  -37687  291378 1302747

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   5882172.9  2480499.9   2.371   0.0215 *
Vol$`Mês/Ano`    -166.9      132.4  -1.261   0.2129
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 461400 on 52 degrees of freedom
Multiple R-squared:  0.02967,     Adjusted R-squared:  0.01101
F-statistic:  1.59 on 1 and 52 DF,  p-value: 0.2129
```

Observando a variável “Vol$`Mês/Ano`”, concluímos que ela não é significativa, logo, podemos concluir que a nossa série continua estacionária.

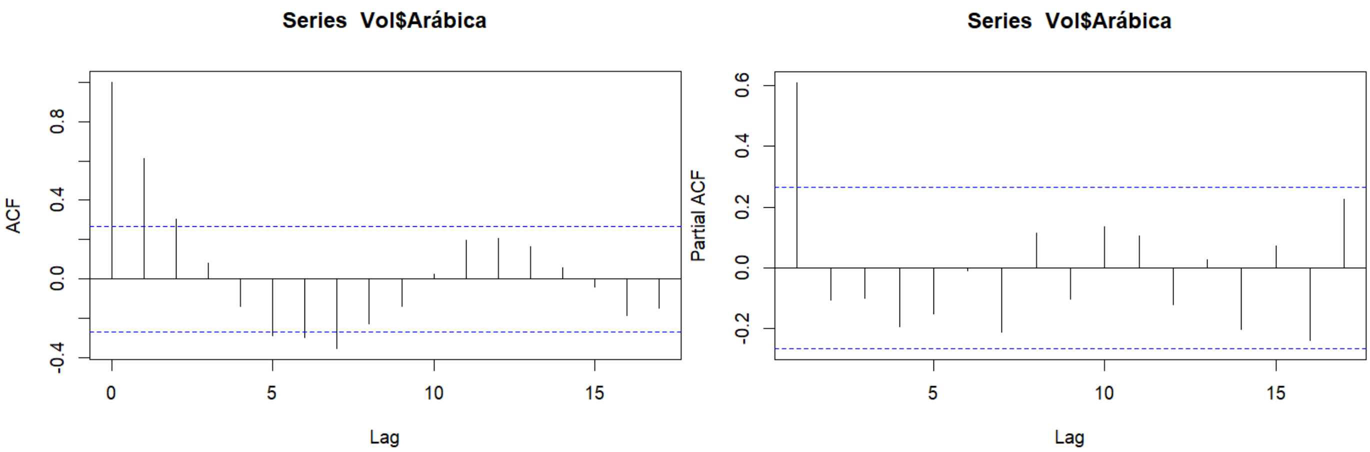
PACF, ACF e Modelos

Como ela é estacionária, podemos observar outras informações da série: O ACF e o PACF.

A Função de Autocorrelação (ACF) mede a correlação entre os valores da série temporal em diferentes lags, enquanto a Função de Autocorrelação Parcial (PACF) mede a correlação entre os valores da série temporal em diferentes lags, mas removendo o efeito das autocorrelações nos lags intermediários.

Sendo assim, o ACF e o PACF ajudam a descobrir modelos de séries temporais. Observando os gráficos:

Figura 3: Gráfico do ACF e PACF da série



Observando os gráficos, podemos perceber no ACF que 2 (até mesmo 3 lags) são significativos, enquanto no PACF, apenas 1 é significativo, dando a ideia de um modelo ARIMA(1,0,1), ARIMA(1,0,2), ARIMA(1,0,3), ou até mesmo, pelo fato de alguns lags do PACF estarem próximos de serem significativos, arriscaria um ARIMA(2,0,1). (modelos ARMA).

Utilizando a função “auto.arima()” do programa R, obtemos uma recomendação de modelo. Logo:

```
> model_arima_100 = auto.arima(Vol$Arábica)
> model_arima_100
```

Series: Vol$Arábica

ARIMA(1,0,0) with non-zero mean

Coefficients:

ar1 mean

0.6315 2743985

s.e. 0.1064 128294

sigma^2 = 1.331e+11: log likelihood = -767.45

AIC=1540.91 AICc=1541.39 BIC=1546.88

O modelo recomendado para nós, é o ARIMA(1,0,0) (modelo AR). O segundo modelo que irei escolher, dos que eu identifiquei, será o ARIMA(2,0,1).

Observando os coeficientes do modelo ARIMA(1,0,0):

Tabela 2: Coeficiente e p-valor do modelo ARIMA(1,0,0)

Modelo	**Ar1**
ARIMA(1,0,0)	0.6315
P-value	2.897e-09

Sendo nosso ruído branco: sigma^2 = 1.331e+11

Tendo um alfa de 0.05, podemos dizer que o coeficiente estimado é significativo!

Agora, para o ARIMA(2,0,1):

Tabela 3: Coeficientes e p-valores do modelo ARIMA(2,0,1)

Modelo	**Ar1**	**Ar2**	**Ma1**
ARIMA(2,0,1)	1.5755	-0.6826	-1
P-value	2.2e-16	2.224e-11	2.2e-16

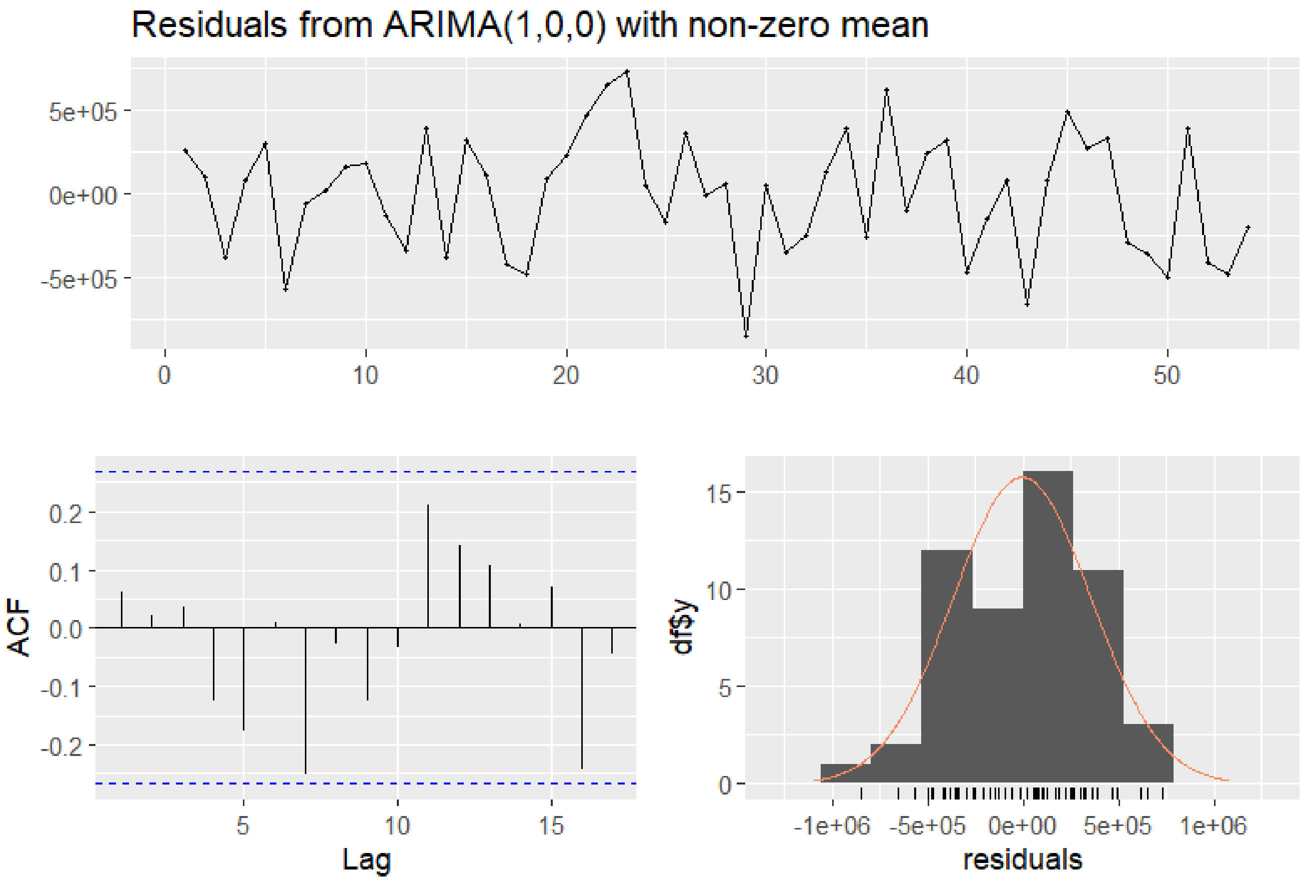
Sendo nosso ruído branco: sigma^2 = 1.21e+11

Tendo um alfa de 0.05, podemos dizer que os coeficientes estimados são significativos!

Análise Residual

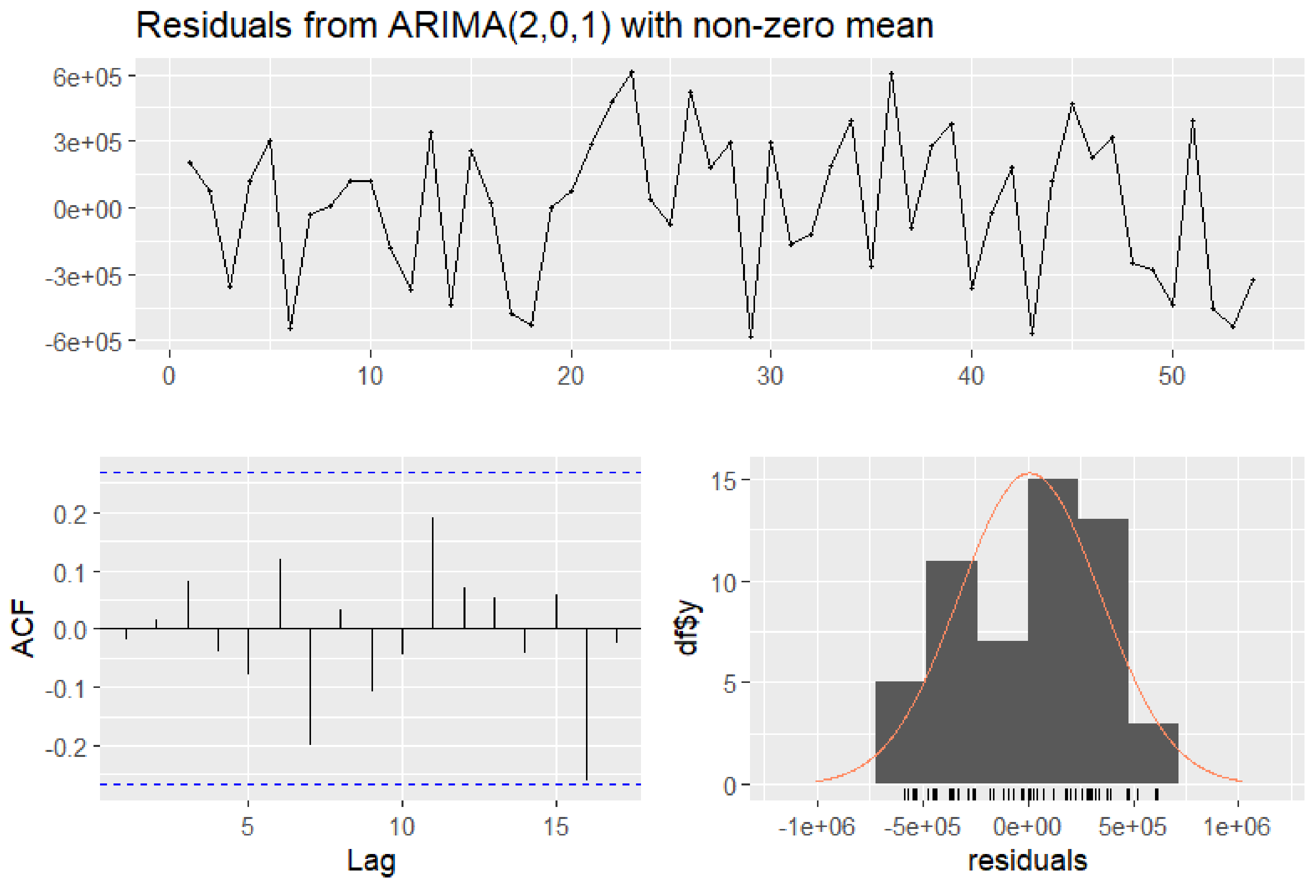
Vamos fazer a análise residual dos modelos escolhidos. Vamos fazer os gráficos de pontos e seus histogramas:

Figura 4: Gráfico dos resíduos do modelo ARIMA (1,0,0)



Vemos que para o modelo ARIMA(1,0,0), possa ter uma normalidade, se vermos o histograma do gráfico.

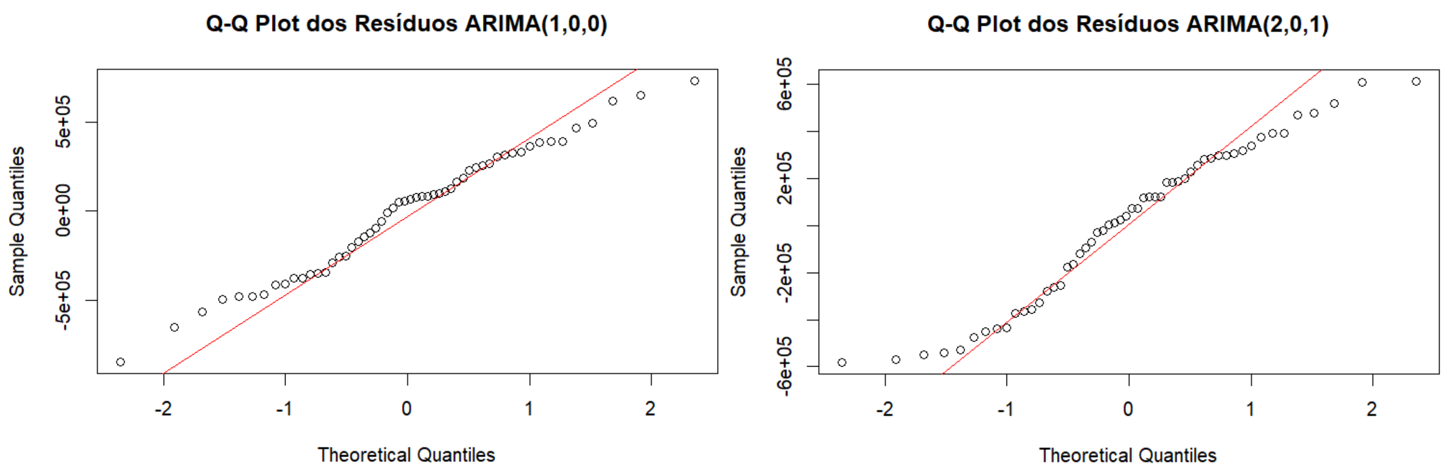
Figura 5: Gráfico dos resíduos do modelo ARIMA (2,0,1)



Para o ARIMA(2,0,1) também há uma certa normalidade nos dados, se observarmos o histograma dos resíduos.

Fazendo os gráficos probabilístico para os dois modelos, também conseguimos observar uma certa normalidade nos resíduos:

Figura 6: Gráfico dos resíduos para verificar normalidade



Para termos uma melhor confirmação, vamos fazer os testes de hipóteses para verificar normalidade nos resíduos. Vamos utilizar o Shapiro-Wilk e do Jarque-Bera. Os dois testes tem praticamente as mesmas hipóteses, porém o de Jarque-Bera é um teste específico para séries:

H0: Os dados seguem uma distribuição normal.

H1: Os dados não seguem uma distribuição normal.

Rejeitamos H0 se o P-valor for menor que o nível de significância. Aplicando para os dois modelos, temos:

```
> shapiro.test(res100)

Shapiro-Wilk normality test

data:  res100
W = 0.98168, p-value = 0.575

> shapiro.test(res201)

Shapiro-Wilk normality test

data:  res201
W = 0.95716, p-value = 0.05147

> jarque.bera.test(res100)

Jarque Bera Test

data:  res100
X-squared = 1.1968, df = 2, p-value = 0.5497

> jarque.bera.test(res201)

Jarque Bera Test
```

data: res201

X-squared = 2.7493, df = 2, p-value = 0.2529

Temos então que os p-valores são maiores que o alfa = 0.05, logo, existe uma normalidade nos resíduos.

Agora, vamos verificar a presença de autocorrelação nos resíduos dos modelos testados. Verificar a ausência de autocorrelação nos resíduos é essencial para validar a adequação do modelo ajustado, para isso, iremos utilizar o teste do Box-Ljung-Pierce, sendo suas hipóteses:

H0: Os resíduos são independentes e não auto correlacionados até um determinado lag k

H1: Os resíduos não são independentes e há autocorrelação significativa em pelo menos um dos lags até o lag k.

Se o p-valor for maior que o nível de significância, não rejeitamos H0, indicando que não há autocorrelação significativa nos resíduos (os resíduos seguem um processo de ruído branco), sugerindo que o modelo ajustado capturou bem a estrutura de dependência da série temporal. Aplicando o teste:

```
> Box.test(res100, lag = 36, type = "Ljung-Box")
```

Box-Ljung test

data: res100

X-squared = 47.385, df = 36, p-value = 0.09705

```
> Box.test(res201, lag = 36, type = "Ljung-Box")
```

Box-Ljung test

data: res201

X-squared = 43.754, df = 36, p-value = 0.1755

Como os p-valores são maiores que alfa = 0.05, não rejeitamos H0, indicando que os modelos conseguiram se ajustar bem. Observando os diagnósticos dos modelos, podemos perceber essas confirmações graficamente:

Figura 7: Gráfico do teste de Ljung-Box para o modelo ARIMA (1,0,0)

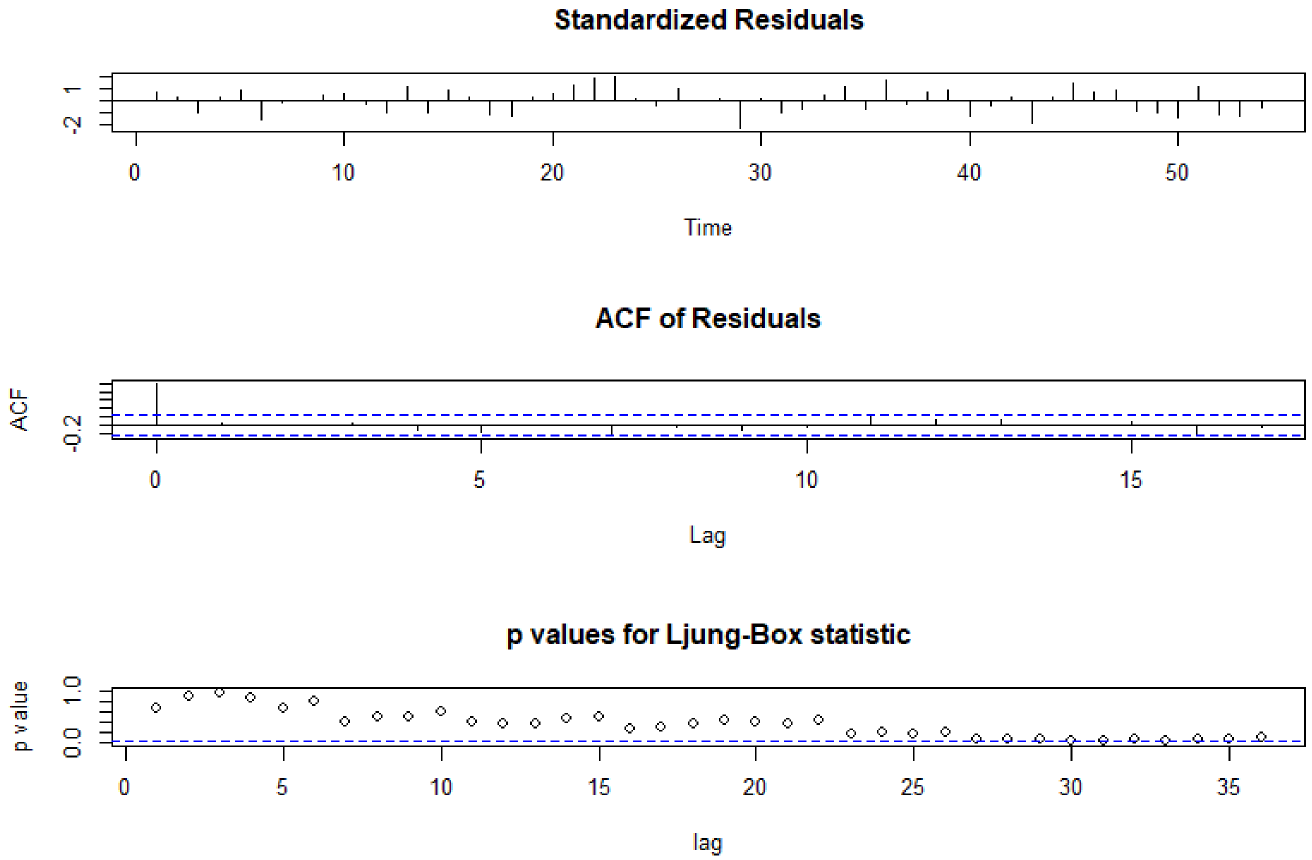
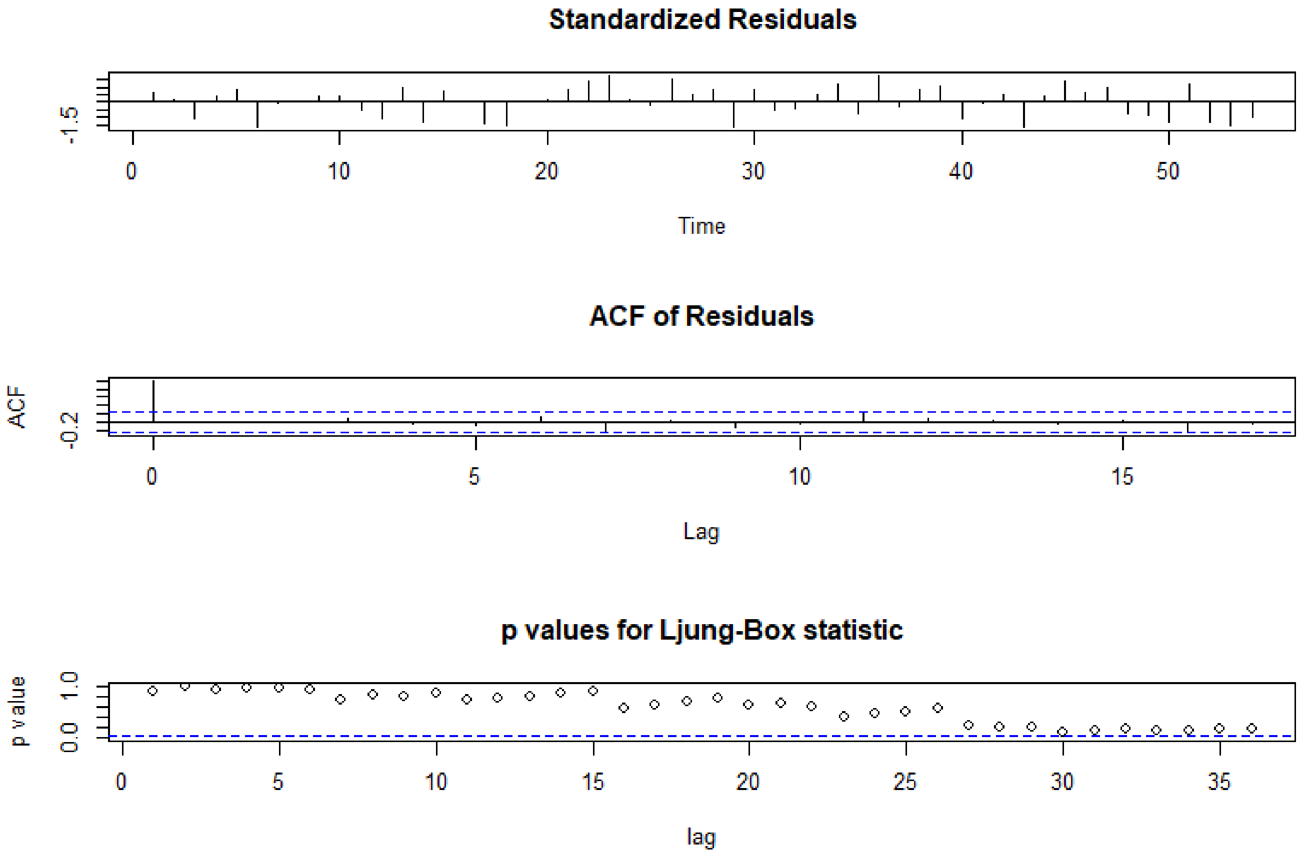


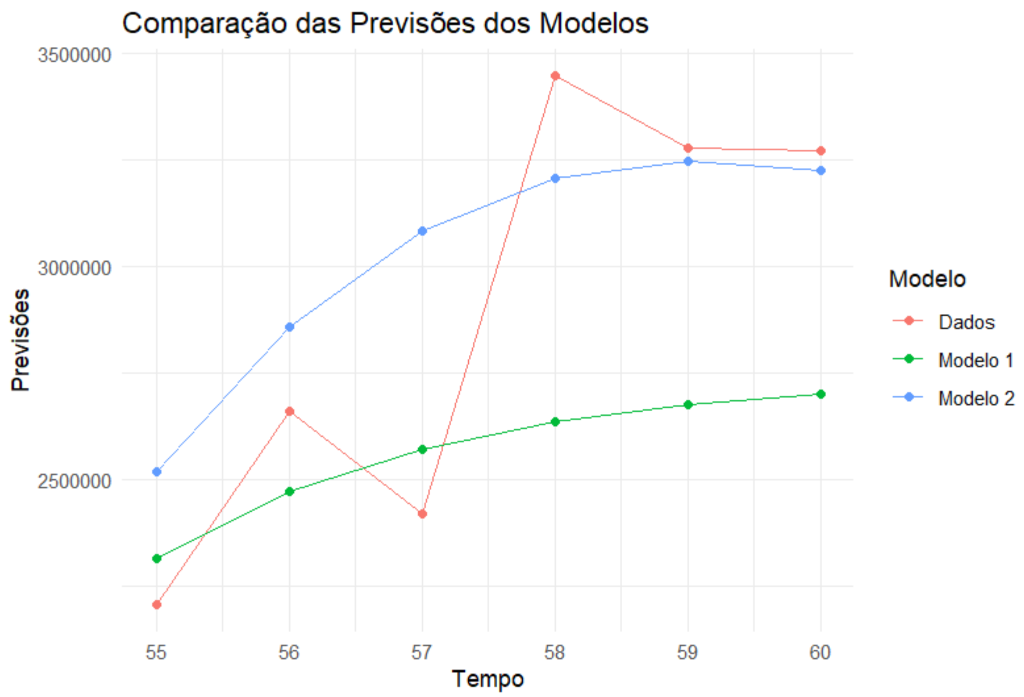
Figura 8: Gráfico do teste de Ljung-Box para o modelo ARIMA (2,0,1)



As Previsões

Verificamos os resíduos, estando tudo ok com os modelos escolhidos. Agora, vamos para a parte das previsões. Com a ajuda do R, temos o gráfico:

Figura 9: Gráfico das previsões dos modelos



Aqui, observamos as previsões dos modelos, sendo “Modelo 1”, o modelo ARIMA(1,0,0) e o “Modelo 2”, o ARIMA(2,0,1). Já os “Dados”, são os nossos verdadeiros valores, aquelas 6 últimas observações que tiramos anteriormente da base de dados. Sendo assim, percebemos que o Modelo 2 está um pouco mais próximo de acertar do que o Modelo 1. Vendo os valores que eles previram, também passa essa sensação:

Tabela 4: Comparação entre as previsões e dados reais

Modelo 1	2312988	2471799	2572092	2635430	2675430	2700690
Dados	2204631	2661527	2419056	3448194	3278414	3273181
Modelos 2	2517029	2857615	3083313	3206442	3246387	3225280

Para conseguirmos ter uma visão melhor dos dados, vamos observar seus intervalos de confiança, para sabermos que estamos errando muito ou pouco:

Tabela 5: Intervalo de confiança para o Modelo 1

Limite Inferior	1597826	1625963	1679461	1724806	1757728	1780181
Modelo 1	2312988	2471799	2572092	2635430	2675430	2700690
Limite Superior	3028150	3317635	3464724	3546054	3593131	3621200

Tabela 6: Intervalo de confiança para o Modelo 2

Limite Inferior	1829463	2058601	2266141	2389505	2419750	2375678
Modelos 2	2517029	2857615	3083313	3206442	3246387	3225280
Limite Superior	3204595	3656630	3900485	4023379	4073024	4074883

Pelos intervalos calculados, podemos observar, menos que seja pouco, que o Modelo 2 esteja errando menos que o Modelo 1, já que se subtrairmos os limites dos intervalos, o intervalo do Modelo 2 sugere menos espaçamento que o do Modelo 1, podendo ser mais assertivo nas previsões futuras.

Para concluir nossa análise, podemos observar nosso AIC’s e BIC’s dos modelos, e principalmente, o EQMMP (Erro Quadrático Médio de Previsão). O

EQMMP é usado para avaliar a precisão das previsões de um modelo de séries temporais. Ele calcula a média dos quadrados das diferenças entre os valores previstos e os valores reais observados. Logo:

Tabela 7: AIC, BIC e o EQMMP dos modelos

	Modelo 1	**Modelo 2**
AIC	1540.909	1540.079
BIC	1546.876	1550.024
EQMMP	237179833180	106507451915

Pelos valores dados, não conseguiríamos nos decidir pelos AIC's e BIC's, já que no Modelo 2, o AIC dele é menor que o do Modelo 1, mas ocorre o contrário se observarmos os BIC's. Com o EQMMP, podemos desempatar, sendo o do Modelo 2 é menor do que o do Modelo 1.

Conclusão

As séries temporais são fundamentais em diversas áreas por sua capacidade de capturar e analisar dados que variam ao longo do tempo. Elas são essenciais para prever tendências e padrões, permitindo que empresas e pesquisadores façam previsões informadas sobre o futuro com base em dados históricos. Com um bom modelo, podemos observar um futuro decaimento ou aumento, e dependendo do objetivo, podemos tentar evitar esse futuro decaimento ou aumento.

Em relação do nosso estudo sobre a exportação do café arábica, quanto mais exportamos, melhor para nós, pois geraria mais lucro para o Brasil. Sendo assim, um bom modelo poderia prever esses decaimentos e tentar criar uma estratégia para evitar isso, já que o cultivo do café precisa ter muito cuidado, e perder safras de café por não estar sendo tão exportado é muita perda de lucro.

Sendo assim, o nosso objetivo foi tentar criar um modelo que conseguisse capitar, pelo menos esse decaimentos e aumentos no futuro, e com base no que analisamos, o Modelo 2 (ARIMA(2,0,1)) se mostrou muito bem para o caso, já que o nosso critério de desempate foi o EQMMP, e o do ARIMA(2,0,1) foi menor do que o do outro modelo, concluindo que ele era melhor que o do ARIMA(1,0,0). Vale ressaltar que no trabalho feito, analisamos outros modelos (modelo ARIMA(1,0,1); ARIMA(1,0,2); ARIMA(1,0,3); ARIMA(1,0,4); ARIMA(1,0,5) e ARIMA(1,0,6)), porém, eles não passaram no teste dos coeficientes, já que alguns coeficientes não foram significativos. Entretanto, fiz a análise completa deles, porém, no final, a melhor escolha ainda foi o modelo ARIMA(2,0,1), então, para não ficar tanta coisa, escolhi fazer o relatório com os 2 melhores modelos que observei (ARIMA (1,0,0) e ARIMA(2,0,1)).

Referências

DOMANI CONSULTORIA. Exportação de café. Disponível em: https://www.domaniconsultoria.com/post/exporta%C3%A7%C3%A3o-de-caf%C3%A9. Acesso em: 09 jun. 2024.

CECAFÉ. Relatório de exportações. Disponível em: https://www.cecafe.com.br/publicacoes/relatorio-de-exportacoes/. Acesso em: 09 jun. 2024.

CECAFÉ. Brasil exporta 39,2 milhões de sacas de café em 2023. Disponível em: https://www.cecafe.com.br/publicacoes/noticias/brasil-exporta-392-milhoes-de-sacas-de-cafe-em-2023-20240115/. Acesso em: 09 jun. 2024.

COOXUPÉ. Afinal, o que é café arábica? Disponível em: https://hubdocafe.cooxupe.com.br/afinal-o-que-e-cafe-arabica/. Acesso em: 09 jun. 2024.

Código do Trabalho

OBS : No código, eu coloquei apenas a análise do Modelo ARIMA(1,0,0) e do ARIMA(2,0,1):

```
# Trabalho de Séries Temporais
# Thaii Céu

#Bibliotecas usadas

library(ggplot2)
library(tseries)
library(forecast)
library(lmtest)

#Séries de Exportação de café (Arábica) : Variável Volume

head(Volume)

Volume$`Mês/Ano` = as.Date(Volume$`Mês/Ano`)

VA = ggplot(Volume, aes(x = `Mês/Ano`, y = `Arábica`)) +
  geom_line() +
  scale_x_date(date_labels = "%b %Y", date_breaks = "6 month") +
  labs(title = "Série Temporal do Volume (Arábica)", x = "Ano", y = "Volume (Arábica)")
+
  theme_minimal()
VA

#Testando se é estacionária pelo teste e também pelo modelo

adf.test(Volume$Arábica) # É Estacionária

model = lm(Volume$Arábica ~ Volume$`Mês/Ano`)
summary(model) # Tambem deu estacionária
```

```
#Arrumando a série para jan2019 A jun2023

Vol <- subset(Volume, `Mês/Ano` >= as.Date("2019-01-01") & `Mês/Ano` <= as.Date("2023-07-01"))

# Criando o gráfico da série de jan2019 até jun2023
VA1 = ggplot(Vol, aes(x = `Mês/Ano`, y = Arábica)) +
  geom_line() +
  scale_x_date(date_labels = "%b %Y", date_breaks = "6 month") +
  labs(title = "Série Temporal de jan2019 a jun2023 (Volume Arábica)", x = "Data", y = "Volume(Arábica)") +
  theme_minimal()

VA1

#Testando se a série continua estácionaria (Por modelo e teste)

adf.test(Vol$Arábica) #Não é estacionária por 0,00552

model1 = lm(Vol$Arábica ~ Vol$`Mês/Ano`)

summary(model1) # Deu estácionaria!!

#Identificando modelos

acf(Vol$Arábica)
pacf(Vol$Arábica)

model_arima_100 = auto.arima(Vol$Arábica)

model_arima_100 # auto.arima
model_arima_201 <- Arima(Vol$Arábica, order=c(2,0,1))
```

```
#Estimativas dos parâmetros dos modelos e Ruidos Brancos

coef100 = model_arima_100$coef
coef201 = model_arima_201$coef

coeftest(model_arima_100)
coeftest(model_arima_201)

# Testes Residuais

checkresiduals(model_arima_100, lag = 36)
checkresiduals(model_arima_201, lag = 36)

qqnorm(res100, main="Q-Q Plot dos Resíduos ARIMA(1,0,0)")
qqline(res100, col="red")

qqnorm(res201, main="Q-Q Plot dos Resíduos ARIMA(2,0,1)")
qqline(res201, col="red")

shapiro.test(res100)
shapiro.test(res201)

jarque.bera.test(res100)
jarque.bera.test(res201)

#Teste do modelo

Box.test(res100, lag = 36, type = "Ljung-Box")
Box.test(res201, lag = 36, type = "Ljung-Box")

tsdiag(model_arima_100, gof.lag = 36)
tsdiag(model_arima_201, gof.lag = 36)
```

```
# Previsões

prev100 <- forecast(model_arima_100, h=6)
plot(prev100, main="Previsão com o Melhor Modelo ARIMA para os Próximos 6 Meses")

prev201 <- forecast(model_arima_201, h=6)
plot(prev201, main="Previsão com o Melhor Modelo ARIMA para os Próximos 6 Meses")

# Criando um gráfico para juntar as previsões e comparar com os dados reais

VA_6valors = tail(Volume$Arábica, 6)
VA_6valors

prev_valor100 = prev100$mean
prev_valor201 = prev201$mean

data_ts <- ts(Vol$Arábica, frequency=12)

forecast_df <- data.frame(
  Time = seq(length(data_ts) + 1, length(data_ts) + 6),
  Modelo = rep(c("Modelo 1", "Modelo 2", "Dados"), each=6),
  Previsao = c(prev_valor100, prev_valor201, VA_6valors)
)

ggplot(forecast_df, aes(x=Time, y=Previsao, color=Modelo)) +
  geom_line() +
  geom_point() +
  labs(title="Comparação das Previsões dos Modelos", x="Tempo", y="Previsões") +
  theme_minimal()

# Intervalo de Confiança
```

```
LI1 = prev100$lower
LS1 = prev100$upper

LI2 = prev201$lower
LS2 = prev201$upper

#EQMMP, AIC e BIC

erro100 = VA_6valors - prev_valor100
erro201 = VA_6valors - prev_valor201

EQMMP100 = (sum((erro100)^2)) / 6
EQMMP201 = (sum((erro201)^2)) / 6

aic_values <- c(AIC(model_arima_100), AIC(model_arima_201))
bic_values <- c(BIC(model_arima_100), BIC(model_arima_201))
```