Previsão da Receita Semanal para Novembro de 2020

Rafael Setti Riedel Sturaro

15/03/2021

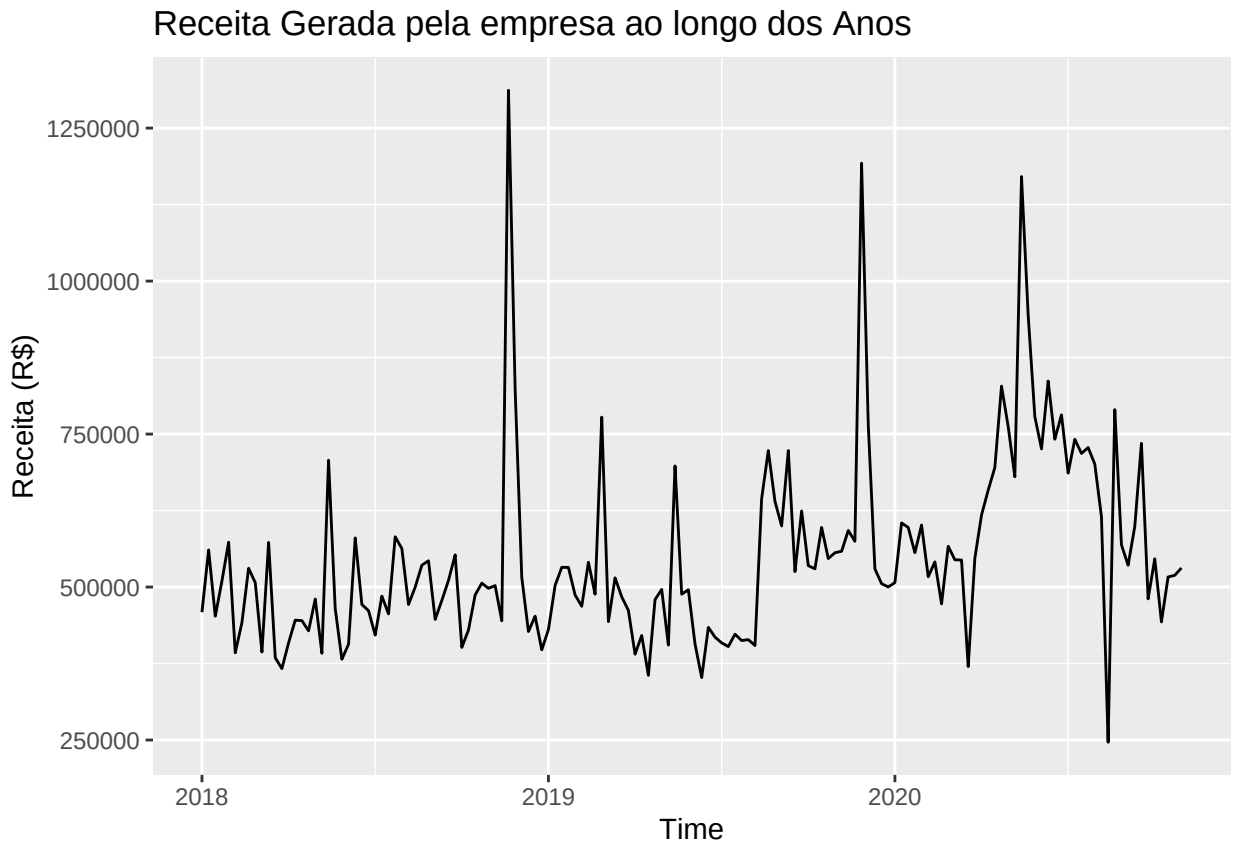
Introdução e Visão Geral do problema

A Black Friday é um momento em que, quando se refere ao ramo de vendas, é sempre esperado uma movimentação grande do mercado. Dessa forma, no contexto do nosso problema, temos um banco de dados que contém a receita de uma empresa, junto a quantidade que o cliente investe em 3 diferentes mídias (A, B e C) coletados semanalmente.

Assim, as mídias A, B e C modelam o valor de nossa variável de interesse (Receita), de forma que, serão propostos modelos de previsão baseados em regressão linear múltipla e previsões baseadas em modelos auto-regressivos.

Dessa forma, parte-se para descrição de uma série temporal com cada unidade de tempo referente a uma semana que contém uma observação:

```
## Definição da série
y <- ts(data = as.numeric(dados$Receita), start = 2018, frequency = 52)
autoplot(y) + ggtitle("Receita Gerada pela empresa ao longo dos Anos") + ylab("Receita (R$)")
```



Olhando para essa série temporal do valor da receita, pode-se identificar alguns padrões além de sugerir perguntas a respeito do comportamento da série, dentre essas questões, tem-se:

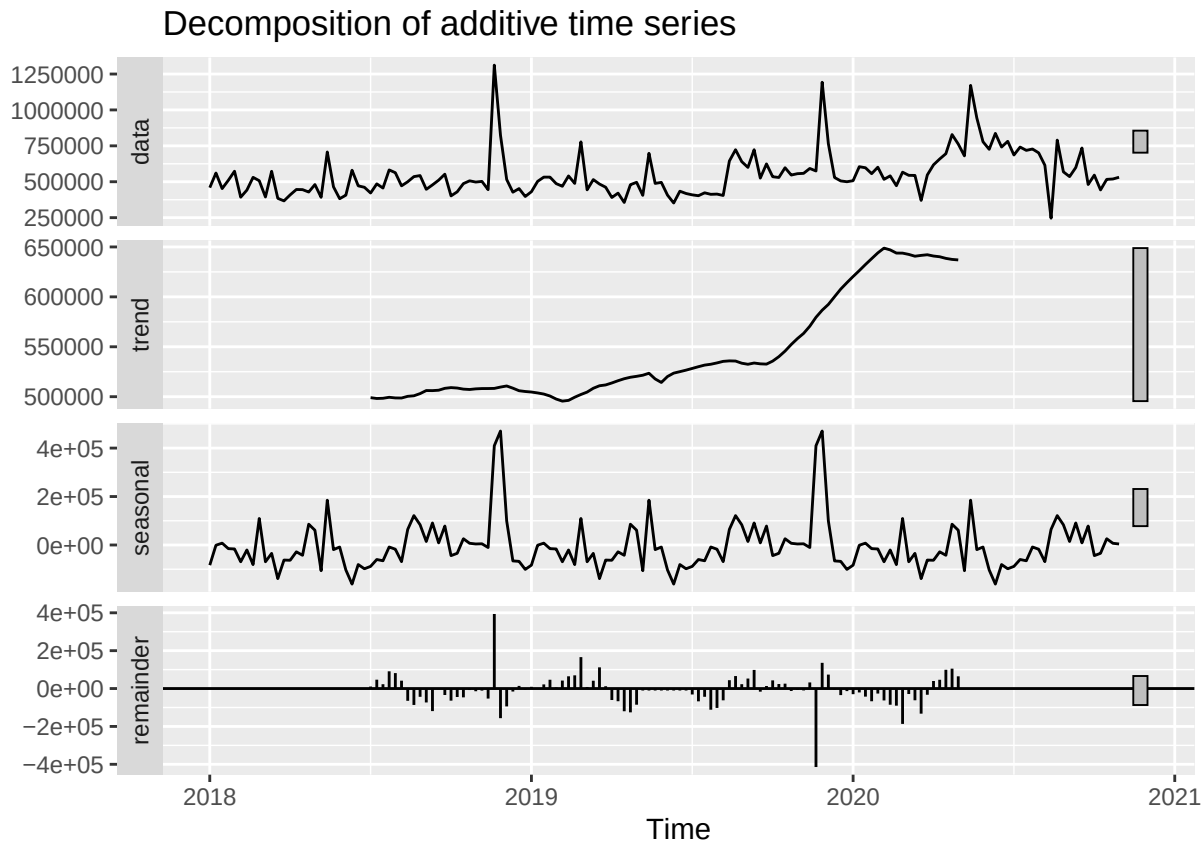
- Dentre esses padrões, primeiramente é sugerido a presença de um período sazonal bastante acentuado, referente justamente a época da Black Friday;
- É possível visualizar, em 2020, uma queda acentuada no valor da receita, seguido por um grande aumento que não foi visualizado nos anos anteriores. Dessa forma, por hora atribuí-se a causa desse evento ao surgimento da pandemia do novo coronavírus, que por tratar de um evento que estimula o funcionamento do e-commerce, pode ser uma das causas desse brusco aumento na receita para o primeiro semestre de 2020;
- Ao longo da coleta dos dados, é possível suspeitar da existência de uma tendência de alta presente na série. Contudo, essa tendência pode ser influenciada pela inflação. Nesse caso, será suposto que a inflação **não** é capaz de influenciar a tendência.

Início do procedimento de previsão

Decomposição da série

A fim de comprovar as suspeitas levantadas no primeiro tópico, será realizada a decomposição da série a fim de identificar tendências e sazonalidades:

```
autoplot(decompose(y))
```

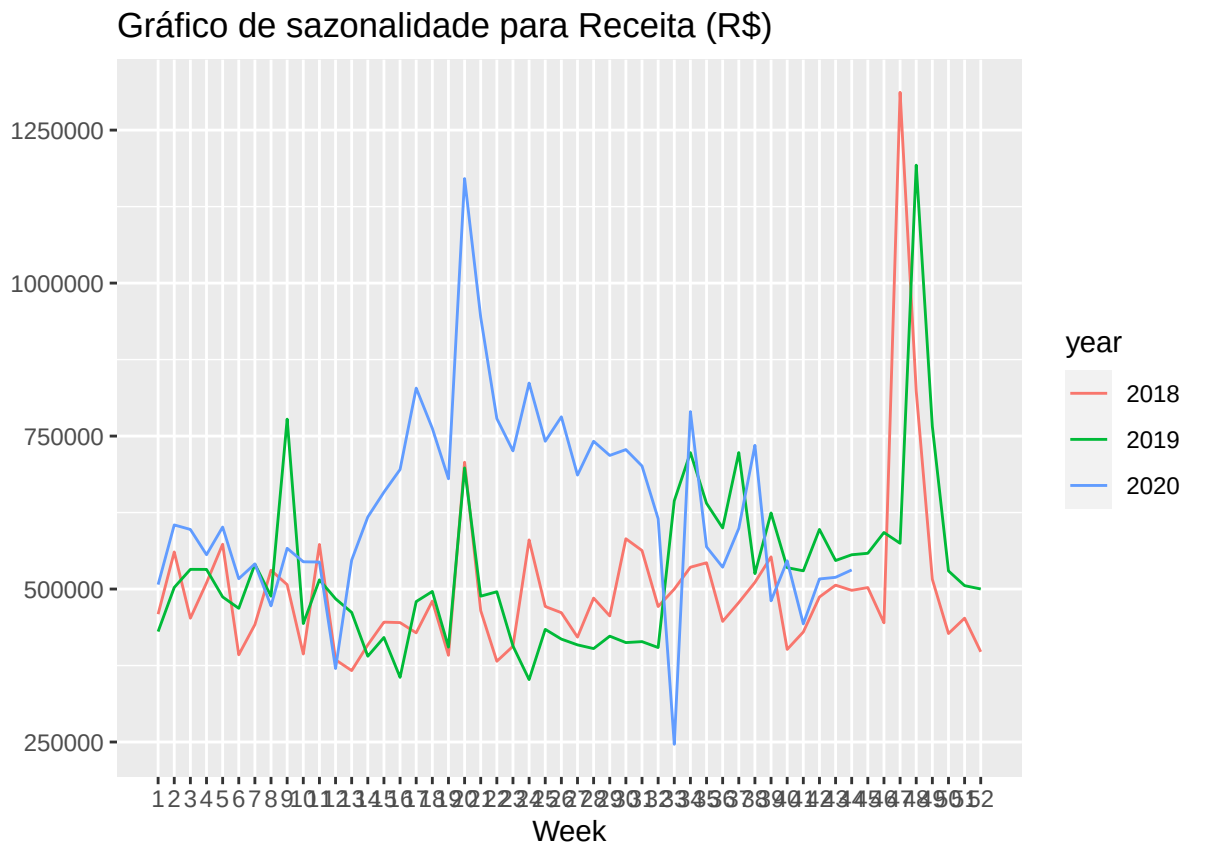


A partir da decomposição da série, é possível identificar que existe, de fato, a tendência de alta que foi proposta anteriormente, tal como a existência de um período sazonal por ano, que é justamente referente a Black Friday. Dessa forma, o modelo de decomposição foi capaz de identificar o período de oscilação presente no surgimento da pandemia do coronavírus como um comportamento anômalo na série.

Gráficos de sazonalidade

A fim de confirmar novamente a presença de um período sazonal presente no modelo, e, nos prognosticar a respeito da previsão exigida, faz-se um gráfico de sazonalidade:

```
ggseasonplot(y) + ggtitle("Gráfico de sazonalidade para Receita (R$)")
```



A partir do gráfico de sazonalidade, dividido pelo período para cada uma da coleta dos dados (Dados semanais), é possível enxergar que nosso período sazonal é justamente as próximas semanas de novembro as quais há interesse em se prever. Assim, pensando nas expectativas para o modelo final de previsão, espera-se um acentuado aumento na receita para esse período a ser previsto.

Outro ponto a ser notado nesse gráfico a disparidade de 1 semana em relação ao pico de venda observado em 2018 e em 2019, fato que pode ser explicado pela Black Friday de 2018 ter ocorrido no dia 23 de novembro (penúltima semana do mês), enquanto a promoção em 2019 ocorreu no dia 29 de novembro (última semana do mês), fato que também é consequência do número de semanas não ser fixo por ano, sendo possível observar a presença de 53 semanas em 2019 e 52 semanas em 2018.

Modelos de séries temporais baseados em regressão múltipla

Validação Cruzada

Para validar a estimação proposta pelos modelos, primeiramente é necessário dividir os dados de interesse em 2 partições: Partição de treino e Partição de teste e, então, comparar a estimação dos parâmetros dada através da partição de treino e comparar com os valores presentes na partição de teste;

```
training <- window(y, start = c(2018,01), end = c(2020,13))
testing <- window(y, start = c(2020,14))
```

Modelo de regressão linear múltipla aproximado por pares de Fourier

Um primeiro modelo a ser sugerido é construído de forma a assumir que os valores das covariáveis são capazes de explicar os valores futuros da variável de interesse (Receita).

Para isso, é construído um modelo de regressão linear múltipla incluindo as 3 covariáveis Investimento em Mídias A, B e C) no processo, de forma que descreve-se o modelo como:

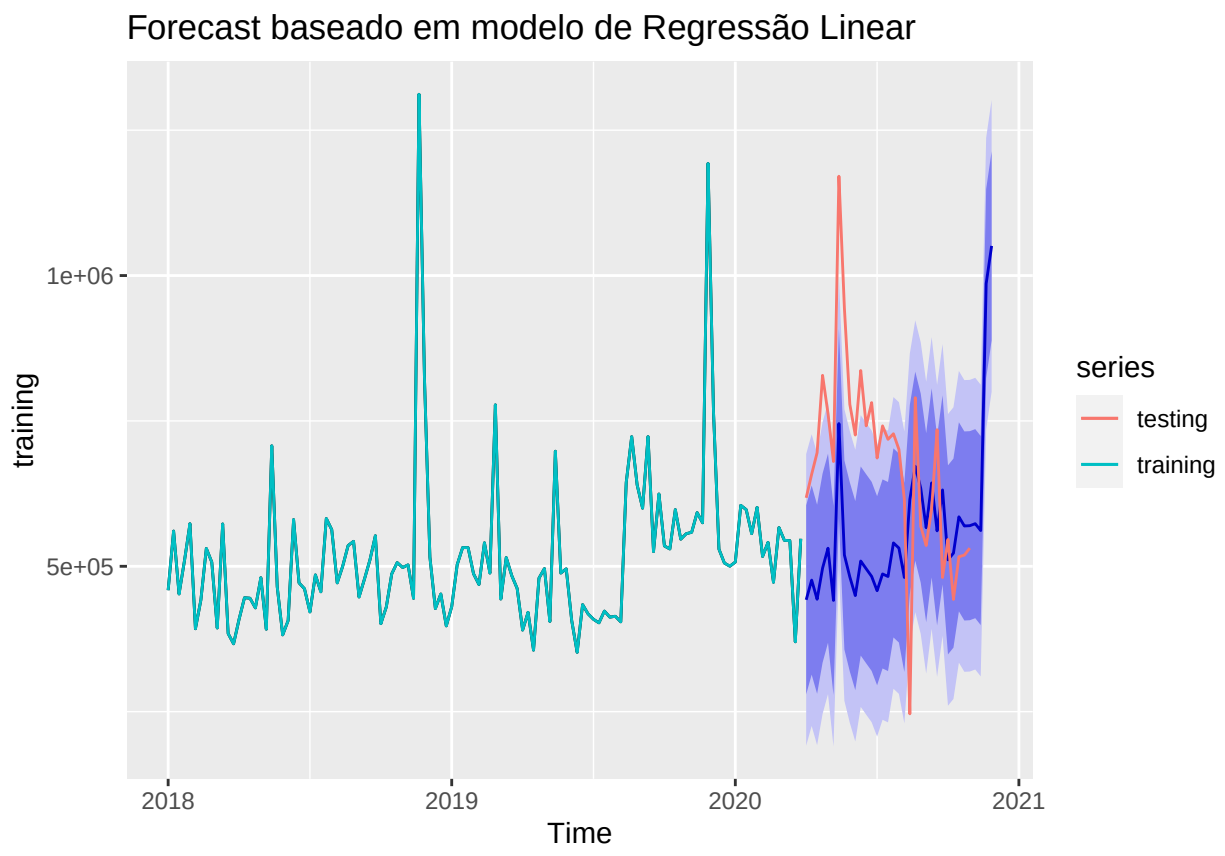
```
mt <- tslm(training~trend+fourier(training,26))
```

Como os dados possuem uma frequência muito alta (aproximadamente 52 semanas por ano), uma boa alternativa para lidar com a sazonalidade é utilizar a aproximação de Fourier para uma função periódica modelado pelos pares de Seno e Coseno incluindo as 3 covariáveis.

No nosso caso, o número máximo de termos de Fourier é dado por metade do período da nossa série (Aproximadamente 52).

Assim, obtem-se um forecast ilutrado por:

```
mt <- tslm(training~trend+fourier(training,26))
ft <- forecast(mt, data.frame(fourier(training,26,length(testing)+4)))
autoplot(ft) + autolayer(training) + autolayer(testing) +
  ggtitle("Forecast baseado em modelo de Regressão Linear")
```



Conclui-se, então, a repeito do modelo de regressão múltipla, que este foi capaz de ajustar os dados com certa precisão, levando em conta o pequeno número de observações presente para estudo.

O modelo foi capaz de prever uma alta relacionada ao período anômalo ao longo do primeiro semestre de 2020 (Atribuído ao surgimento da pandemia da COVID-19).

Por fim, observa-se que o modelo prevê uma alta na receita para as próximas semans de novembro, periodo sazonal em que estamos interessados (Black Friday), de forma que pode-se esperar um pico de vendas semelhante ao de 2019.

Vale também ressaltar que, para esse gráfico, as escalas estão ajustadas devido ao comportamento de tendência presente na série original.

Modelos Auto-Regressivos

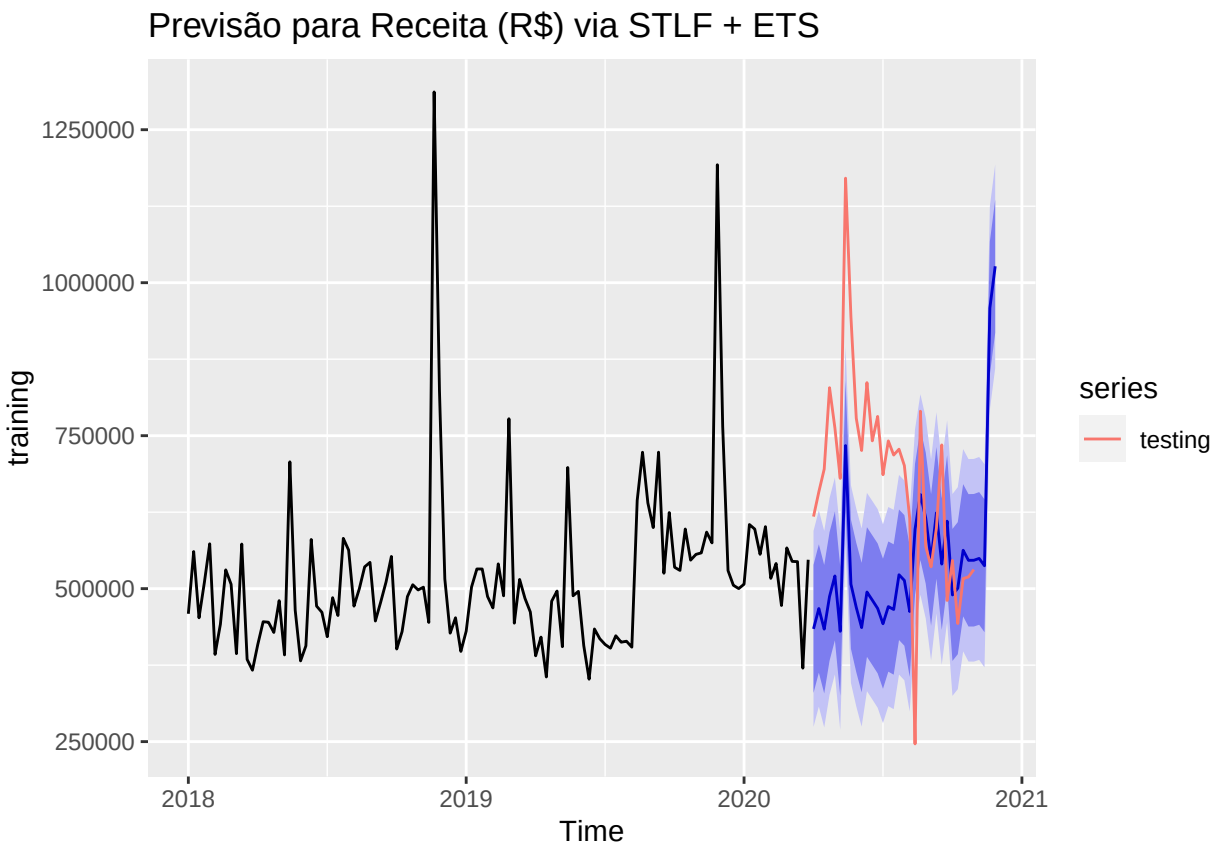
Modelo STLF

Como potencial primeiro modelo auto-regressivo, tem-se o modelo STLF que se dá combinado ao modelo de alisamento exponencial para realizar o primeiro forecast auto-regressivo. Descreve-se o modelo como:

```
m1 <- stlf(training)
```

Que resultou em uma previsão ilustrada por:

```
f1 <- forecast(m1, h = length(testing) + 4)
autoplot(f1) + autolayer(testing) + ggtitle("Previsão para Receita (R$) via STLF + ETS")
```



É possível notar uma disparidade dos valores reais com os valores previstos pelo modelo quando comparados, principalmente ao período do surgimento da COVID-19, o que causou um impacto que o modelo não foi capaz de representar.

Quanto as próximas semanas da previsão (Referentes ao mês de novembro e a própria Black Friday), pode-se notar que esse modelo é capaz de identificar o período sazonal da Black Friday, e então, se mostra eficiente em ao menos indicar um forte indicador de que a receita deve subir ao longo do mês de Novembro.

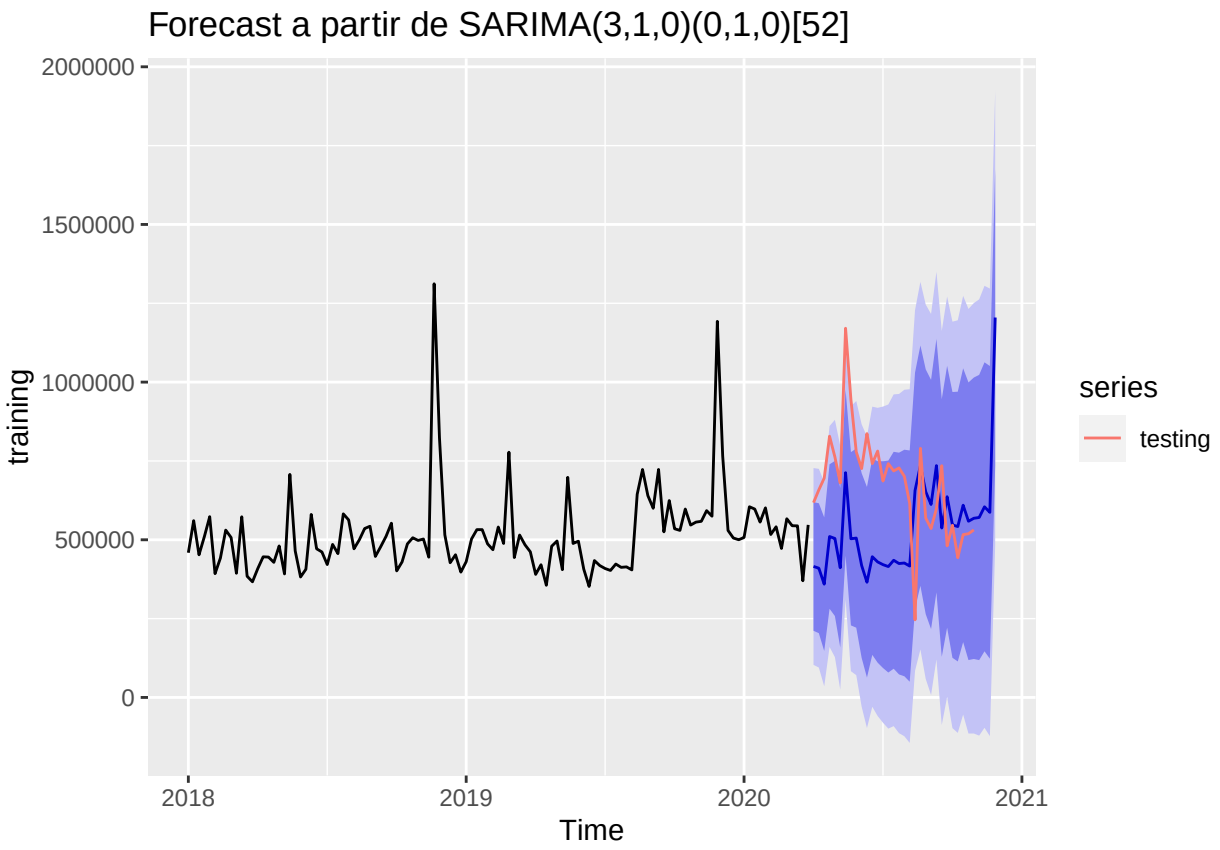
Modelo ARIMA

O segundo modelo a ser desenvolvido é um modelo ARIMA. Contudo, como a série possuí tendência e sazonalidade, é necessário aplicar diferenciação nos dados a fim de transformar a série em estacionária e, adicionar o efeito sazonalidade ao modelo. Transformando-o em, na verdade, um modelo SARIMA.

```
m2 <- auto.arima(training, d=1, D=1, stepwise = FALSE)
```

Que resultou em um forecast ilustrado por:

```
f2 <- forecast(m2, h = length(testing)+4)
autoplot(f2) + autolayer(testing) + ggtitle("Forecast a partir de SARIMA(3,1,0)(0,1,0)[52]")
```



A partir do gráfico, nota-se que reais valores da partição de teste estão contidos dentro do intervalo de confiança para os valores ajustados, porém ainda com uma certa defasagem.

Para os valores a serem preditos no mês de novembro, novamente o modelo foi capaz de identificar o período sazonal referente a Black Friday e então, retornou novamente uma grande receita para as próximas semanas.

Modelo de Regressão Harmônica Dinâmica

Como último candidato a modelo, será feito um modelo de Regressão Harmônica Dinâmica, de forma que é selecionado um par de coeficientes para a função ARIMA de forma a minimizar o AICc.

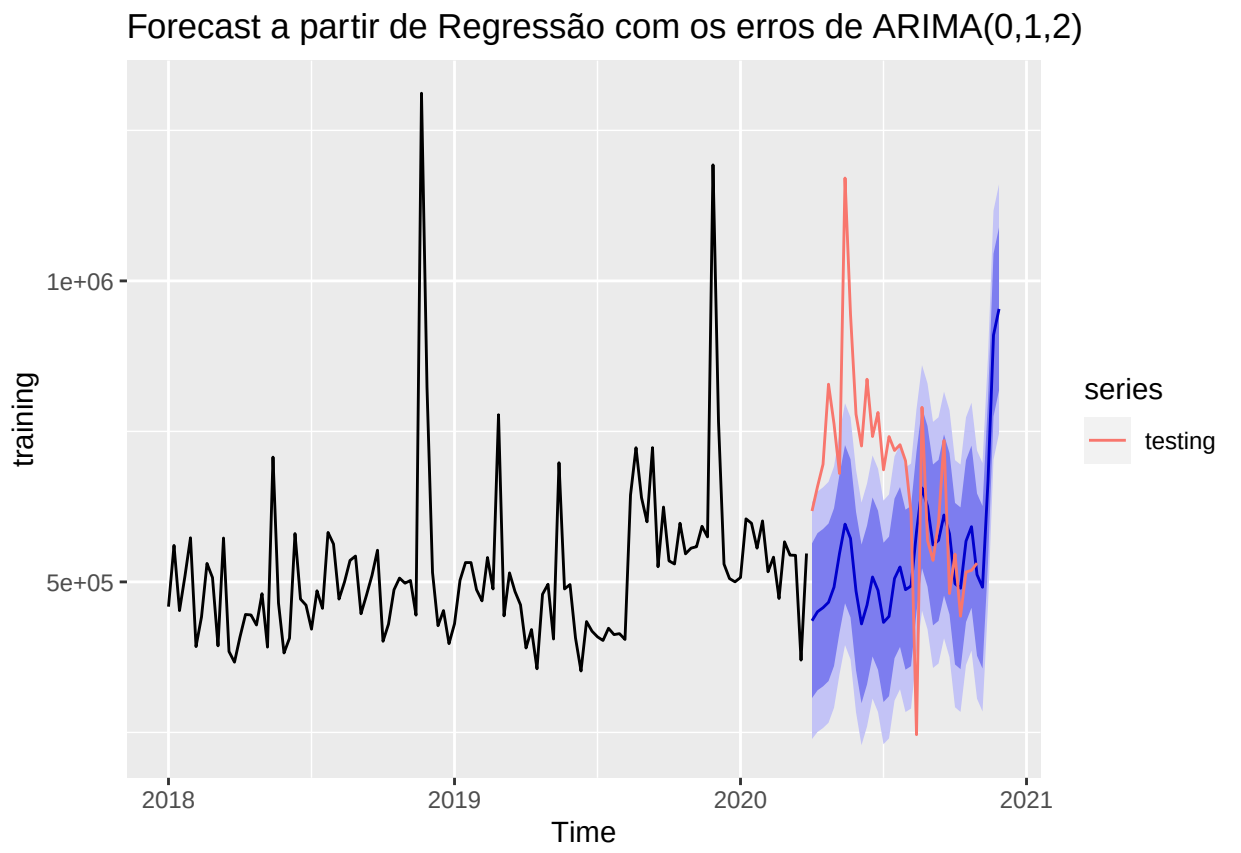
Esse modelo foi sugerido por se tratar de dados colhidos semanalmente, e as previsões realizadas a partir de modelos ARIMA podem acarretar em falhas de aproximação devido ao longo período (Aproximadamente 52). Além de poder ser aplicado a séries com tendência e sazonalidade.

```
bestfit <- list(aicc=Inf)
for(K in seq(25)) {
  fit <- auto.arima(training, xreg=fourier(training, K=K),
                    seasonal=TRUE)
  if(fit[["aicc"]] < bestfit[["aicc"]]) {
    bestfit <- fit
    bestK <- K
  }
}
```

Que resulta em uma previsão ilustrada por:

```
f5 <- forecast(bestfit,
               xreg=fourier(training, K=bestK, h=length(testing)+4))

autoplot(f5) + autolayer(testing) + ggtitle("Forecast a partir de Regressão com os erros de ARIMA(0,1,2
```



Novamente para esse modelo, o período classificado como o surgimento da pandemia da COVID-19 não pode ser explicado. De forma que os valores presentes na partição de teste não estão contidos em boa parte dos intervalos de confiança sugeridos.

Contudo, novamente o modelo foi capaz de identificar a sazonalidade e retribuir prognósticos de alta na receita para as próximas 4 semanas de novembro.

Modelo Final

Para selecionar o modelo final, é necessário se atentar a algumas métricas que quantificam a qualidade dos modelos propostos. São eles os índices de acurácia RSME, MAPE, MAE e MASE.

Fazendo uma tabela desses coeficientes de acordo com o modelo proposto, visualizamos a seguinte tabela:

	RSME	MAPE	MAE	MASE
Modelo Linear	467678.43	1.336224e+06	467678.43	5.0517472
STL	434240.40	1.240687e+06	434240.40	4.6905578
ARIMA	415434.6	1.186956e+06	415434.64	4.4874226
Regressão Harmônica	434779.21	1.242226e+06	434779.21	4.6963779

A partir da tabela, é possível visualizar que os valores não se diferem muito entre eles, contudo a modelo que apresentou os melhores resultados, dado a tabela, foi o modelo ARIMA.

Contudo, ao longo do procedimento é estudado quanto aos resíduos propostos pelos modelos e, ao visualizar a distribuição dos resíduos obtidos através da ARIMA fez com que fosse descartado como potencial modelo final.

Dessa forma, restaram 3 modelos que a serem considerados, dentre eles um modelo STL, Regressão Harmônica e Modelo Linear.

Por fim, será conduzido o teste de Diebold-Mariano a fim de buscar pelo teste que performa sob maior acurácia.

Após combinar todos os testes entre os modelos restantes, restaram dois modelos a serem considerados: Modelo de previsão STL e modelo de previsão via Modelo Linear. A estatística do teste é dada por:

```
dm.test(f1$residuals,ft$residuals,alternative = "greater",h=30)
```

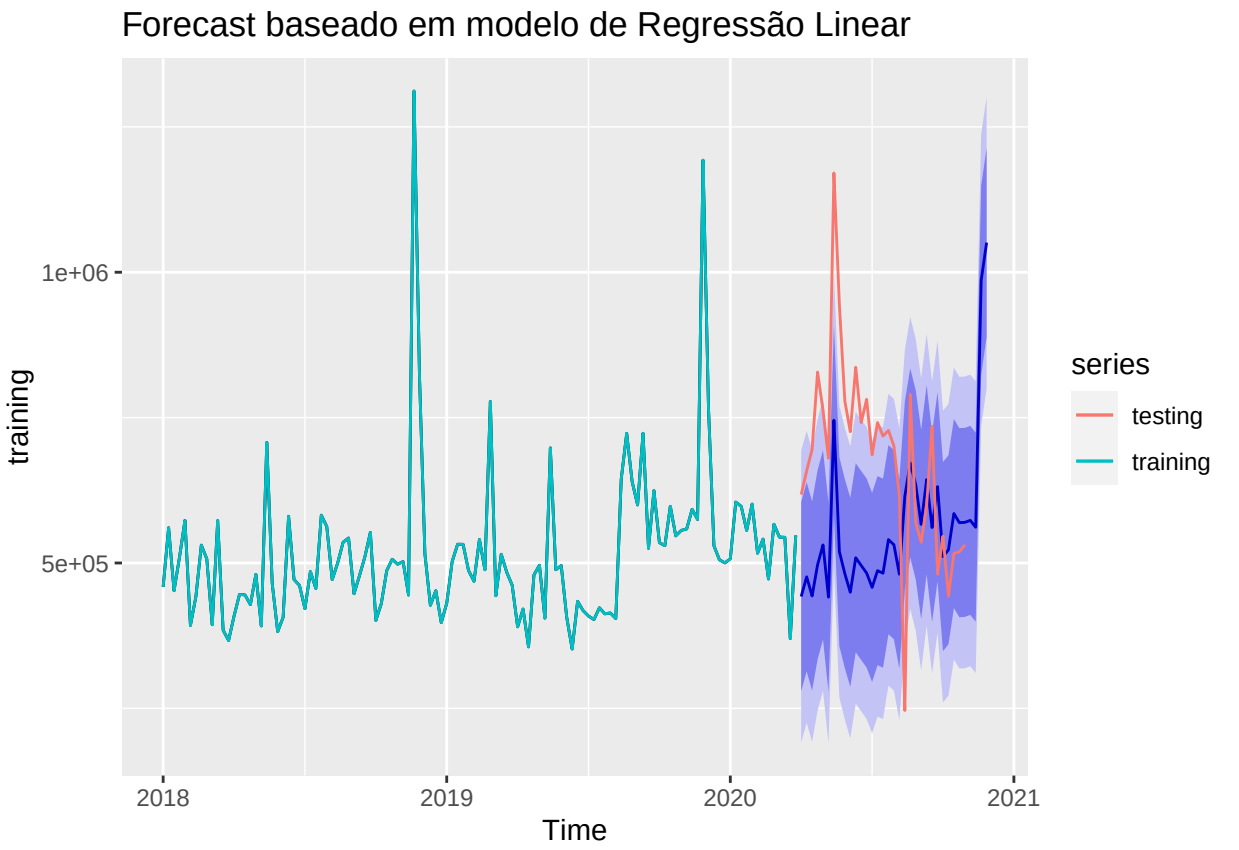
```
##
##  Diebold-Mariano Test
##
## data:  f1$residualsft$residuals
## DM = 2.1196, Forecast horizon = 30, Loss function power = 2, p-value =
## 0.01809
## alternative hypothesis: greater
```

O teste nos retorna a estatística de comparação entre os modelos STL e Linear, e, a partir do teste, como a hipótese nula é rejeitada, toma-se então o modelo previsto via **Regressão Linear com aproximações de Fourier** como o mais preciso e de acordo com as restrições a serem seguidas ao implementar o modelo.

Conclusões a Respeito do Modelo

Relembrando o modelo final dado por:

```
autoplot(ft) + autolayer(training) + autolayer(testing) +
  ggtitle("Forecast baseado em modelo de Regressão Linear")
```



Concluido a respeito do modelo final selecionado, tem-se as seguintes observações finais pautadas:

- Espera-se um crescimento na receita relativo ao período da Black Friday de 2019;
- O modelo se mostrou mais robusto ao se tratar de valores outliers, de forma que soube contornar o período anômalo visualizado no surgimento da pandemia;
- Os valores estão transformados em relação aos valores da série original. Isso se da devido ao método para tratar com a tendência presente na série da receita;
- O investimento em mídias A, B e C são capazes de modelar a receita e agregam ao modelo preditivo.
- Embora o modelo não tenha obtido resultados tão significativamente melhores quando comparados aos outros é interessante ressaltar que, a capacidade computacional exigida para a execução desse modelo foi inferior aos demais, de forma que, ao se tratar de um conjunto de dados maior que o estudado, esse modelo será capaz de suportar o processamento das previsões.