Universidade de São Paulo
Instituto de Ciências Matemáticas e de Computação

Marketing digital: Modelo de previsão para tráfego orgânico

Mariane Romildo dos Santos

São Carlos - SP
Julho/2020

Resumo

Com a crescente utilização da *internet* para a contratação de serviços e compra de produtos, o *marketing* digital se tornou parte indispensável da estratégia de negócios de uma empresa que deseja ser bem sucedida. O detalhamento do perfil do cliente, demográfico e comportamento *online*, faz com que as estratégias sejam tão assertivas. Cada visita a um determinado site gera uma grande quantidade de variáveis, sendo que milhares de visitas acontecem durante um único dia. Para acompanhar todo esse volume de dados e transformá-los em informações, empresas oferecem produtos que facilitam essa tarefa. Algumas delas oferecem algum tipo de inteligência dos dados. Após perceber problemas em previsões oferecidas por uma dessas plataformas, este trabalho faz a comparação desses valores com os dados reais e procura uma relação entre as séries temporais que traria previsões melhores.

Palavras-chave: *Marketing* digital; *Inbound marketing*; Séries temporais; Previsão.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

No momento do seu surgimento, o *marketing* digital era somente uma forma diferente de se fazer *marketing*, nos últimos anos essa forma se tornou parte essencial da estratégia de negócios de uma empresa (KANNAN e LI, 2017). Esse fato se deve ao uso da internet continuar crescendo em todo o mundo e ter se tornado um dos meios mais importantes para transações de produtos e serviços (LEEFLANG et. al, 2014).

Para compor uma estratégia de *marketing* digital ideal para o seu negócio existem algumas formas. Entre as mais conhecidas estão as chamadas mídias de *performance* e o *inbound marketing*. As mídias de *performance*, ou mídias pagas, são responsáveis por veicular anúncios de clientes que pagaram para que eles fossem mostrados, as empresas interessadas participam de um leilão, no qual não só o maior lance conta mas também a qualidade do *site* para o qual esse anúncio está levando o usuário, para que seja decidido qual anúncio aparecerá e em qual posição. Já o *inbound marketing*, tem como principal característica levar visitantes ao site através de mídias não pagas (mídias orgânicas), como, por exemplo, *blogs*, resultados de busca que não sejam patrocinados, publicações em redes sociais e mensagens por *e-mail*.

Empresas que investem em *marketing* digital possuem diversas vantagens sobre os seus concorrentes como, por exemplo, melhor conhecimento do seu público-alvo, direcionamento de anúncios e campanhas para usuários que demonstraram algum tipo de interesse por algum produto, exposição de anúncios que promovem o conhecimento da marca para além dos seus clientes, maior proximidade com os seus clientes por meio de redes sociais, entre outras vantagens.

Por trás das vantagens do *marketing* digital está também um dos seus maiores desafios: o *Big Data*. O detalhamento de dados sobre comportamento dos usuários nas páginas da *web*, seus comentários em redes sociais e dados demográficos é incrível, mas em contrapartida essa quantidade de dados produzida pelos milhões de usuários não pode ser tratada como uma base de dados comuns. A dimensionalidade traz dificuldades para a coleta, tratamento, armazenamento, compartilhamento, visualização e análise desses dados (LEEFLANG et. al, 2014).

O *Big Data*, que é um desafio para alguns, foi visto como uma oportunidade de negócio para outros, inúmeras empresas oferecem produtos que coletam, tratam, armazenam, possibilitam a extração e visualização dos dados. Algumas empresas oferecem até inteligência dos dados em seus produtos, desde previsões considerando o comportamento conhecido dos dados até análise de sentimentos dos comentários que os consumidores fizeram sobre um determinado serviço. As análises fornecidas por esses produtos são feitas por meio de algoritmos e por isso podem fornecer resultados equivocados.

É na solução desse problema que este trabalho está inserido. Uma plataforma que oferece ferramentas para facilitar o monitoramento de *sites*, por seus donos e administradores, oferece também previsões para algumas métricas de desempenho das páginas. A métrica de interesse aqui é o volume de tráfego orgânico mensal, ou seja, a quantidade de acessos resultantes de mídias não pagas que um *site* teve em um mês. Administradores de um site perceberam que as estimativas para essa métrica fornecidas pela plataforma eram sempre muito maiores do que viria a ser o valor real ao final do mês, o que era um grande problema já que as previsões eram utilizadas para orientar metas e planos de ação da equipe responsável pelo *inbound marketing* do *site*.

Dado o contexto, o objetivo deste trabalho é comparar as previsões com o tráfego orgânico real e utilizar essas previsões infladas para fornecer previsões mais próximas da realidade.

2 Metodologia

Os dados utilizados serão a previsão da plataforma e o valor real do volume de tráfego orgânico, de 2017 a abril de 2019. Por se tratarem de dados que são observados sequencialmente ao longo do tempo, a análise será feita utilizando métodos de séries temporais.

Serão avaliadas as componentes de uma série temporal: tendência, sazonalidade e ciclos.

Uma série temporal tem tendência quando é observado um longo período de crescimento ou decrescimento nos dados, já a presença de sazonalidade é caracterizada pelos crescimentos e decrescimentos da série que sempre ocorrem em um determinado período do ano, mês, semana ou dia, e por fim, ciclo é quando ocorrem oscilações na série repetidamente, ao longo da componente de tendência (HYNDMAN e ATHANASOPOULOS, 2019).

2.1 Previsão

Para a previsão dos dados foram levantados possíveis modelos que se encaixam nas condições do estudo.

2.1.1 Modelo de regressão linear

Os modelos de regressão linear aplicados em séries temporais partem do princípio de que uma série temporal Y, resposta, tem uma relação linear com uma série explicativa X, representada pela equação

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t,$$

onde β_0 representa o valor predito quando $x_t = 0$ e o coeficiente β_1 representa a mudança, em média, na previsão quando x_t aumenta em uma unidade.

Ao utilizar modelos de regressão linear algumas suposições acerca dos erros $(\epsilon_1, ..., \epsilon_t)$ devem ser verificadas:

- os erros não são autocorrelacionados;
- os erros não são relacionados a variável preditora x;
- os erros seguem um distribuição normal com média zero e variância σ^2.

A estimação dos coeficientes β é feita através do método dos mínimos quadrados, ou seja, minimizando a soma dos quadrados dos erros.

2.1.2 ARIMA

Os modelos ARIMA, *Auto-Regressive Integrated Moving Average model*, são modelos lineares para séries temporais. A notação comumente utilizada para designar esses modelos é ARIMA(p,d,q). O parâmetro p é o responsável para a parte autoregressiva do modelo, o parâmetro d pela diferenciação de série original até que ela se torne estacionária e, por fim, o parâmetro q corresponde às médias móveis.

Em um modelo autoregressivo o valor predito é resultado de uma combinação linear dos valores passados da série. O modelo AR(p) pode ser representado pela equação

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + ... + \phi_p y_{t-p} + \epsilon_t,$$

onde ϵ_t é o ruído branco.

Para um modelo de médias móveis o valor predito é uma combinação linear dos erros de previsão passados. O modelo MA(q) pode ser representado pela equação

$$y_t = \epsilon_t + \theta_1\epsilon_{t-1} + \theta_2\epsilon_{t-2} + ... + \theta_q\epsilon_{t-q},$$

Combinando os modelos AR, MA e a diferenciação da série original obtem-se um modelo ARIMA(p,d,q), representado pela equação

$$y'_t = \phi_1 y'_{t-1} + ... + \phi_p y'_{t-p} + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q} + \epsilon_t,$$

onde os valores y'_t correspondem a série diferenciada.

A estimação dos parâmetros ϕ_1,...,ϕ_p,θ_1,...,θ_q é feita, pelo *software R*, pelo método de máxima verossimilhança. Esse método visa encontrar valores para os parâmetros que maximizem a probabilidade de serem obtidos os valores que foram observados.

2.1.3 Modelo de regressão dinâmica

Os modelos de regressão dinâmica combinam fatores que são considerados em modelos de regressão e em modelos de séries temporais, ou seja, combina regressores e componentes de séries temporais. A principal vantagem é que aqui os erros podem ser autocorrelacionados. Um modelo de regressão dinâmica pode ser representado pela equação

$$y_t = \beta' X_t + u_t,$$

onde

$$u_t = \phi_1 u'_{t-1} + ... + \phi_p u'_{t-p} + \epsilon_t + \theta_1\epsilon_{t-1} + ... + \theta_q\epsilon_{t-q}.$$

Na equação anterior, o erro u_t segue um modelo ARIMA e ϵ_t, erro do modelo ARIMA, é esperado que seja um ruído branco.

A estimação dos parâmetros do modelo é feita através da minimização da soma de ϵ_t ao quadrado.

2.1.4 Avaliação das previsões

O critério utilizado para a escolha do melhor modelo de previsão será o erro percentual médio absoluto (MAPE), portanto o melhor modelo será o que apresentar menor MAPE.

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y_t}}{y_t}\right| * 100$$

3 Resultados e discussões

Para todas as etapas da análise foi utilizado o *software R*, desde a extração dos dados até a avaliação dos modelos.

3.1 Análise descritiva

Existe uma relação forte entre a série das previsões e a série de valores reais, como pode ser visto na figura 3.1. Esse é um motivo para supor que previsões melhores podem ser obtidas utilizando a informação da previsão inflada.

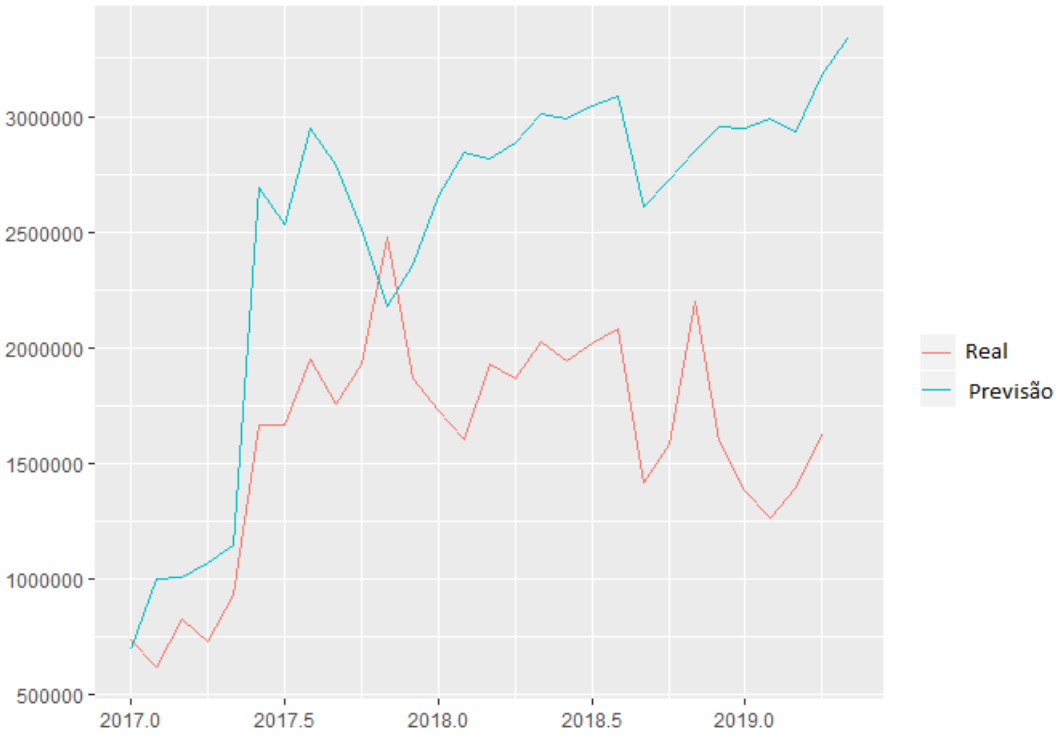

Figura 3.1: Comparação entre a série de valores reais e a série de previsões.

Para confirmar o que foi inferido sobre a partir da figura 3.1 a correlação entre as duas séries temporais foi calculada. O resultado foi uma correlação linear de 0,85, ou seja, uma correlação linear forte (SCHOBER et. al, 2018). Por esse motivo acredita-se que um modelo de regressão linear para séries temporais tenha um bom desempenho (Figura 3.2).

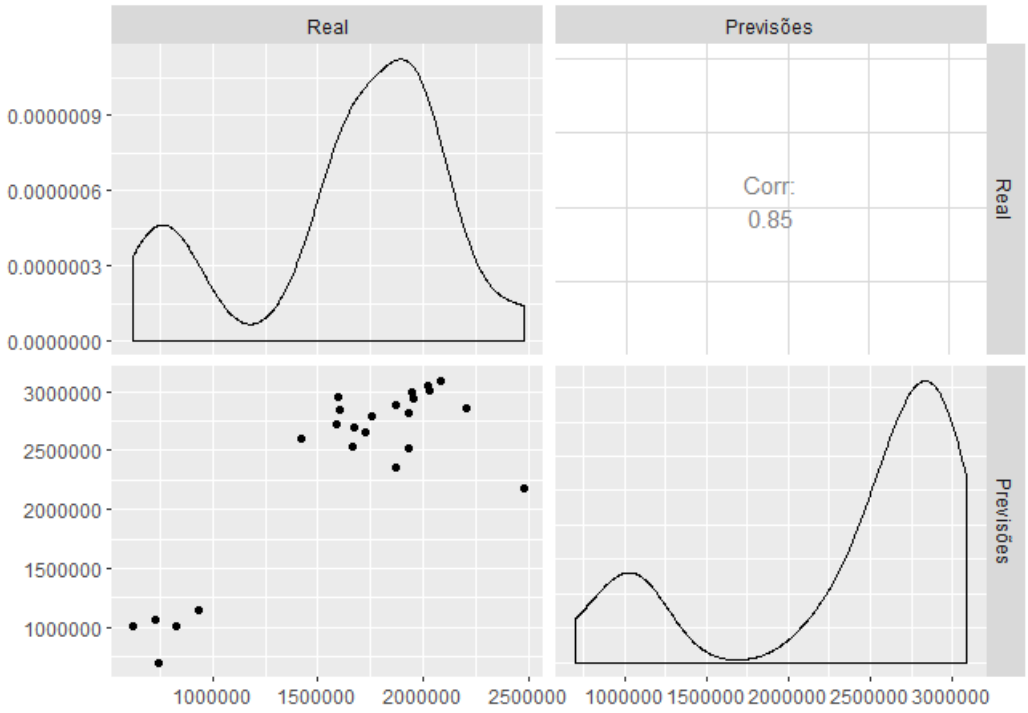

Figura 3.2: Correlação das previsões da plataforma com as observações reais.

Por se tratar de um site de comércio eletrônico, é esperado que o site tenha mais acesso em alguns meses do que em outros por conta de datas comemorativas ou grandes promoções. Essa componente foi avaliada na figura 3.3 e foi verificada uma forte sazonalidade principalmente no comportamento do primeiro trimestre, que se repetiu nos três anos, e no comportamento do último quadrimestre, que se repetiu em 2017 e 2018. A hipotese de que a série tem tendência foi rejeitada pelo teste de *Mann-Kendall*.

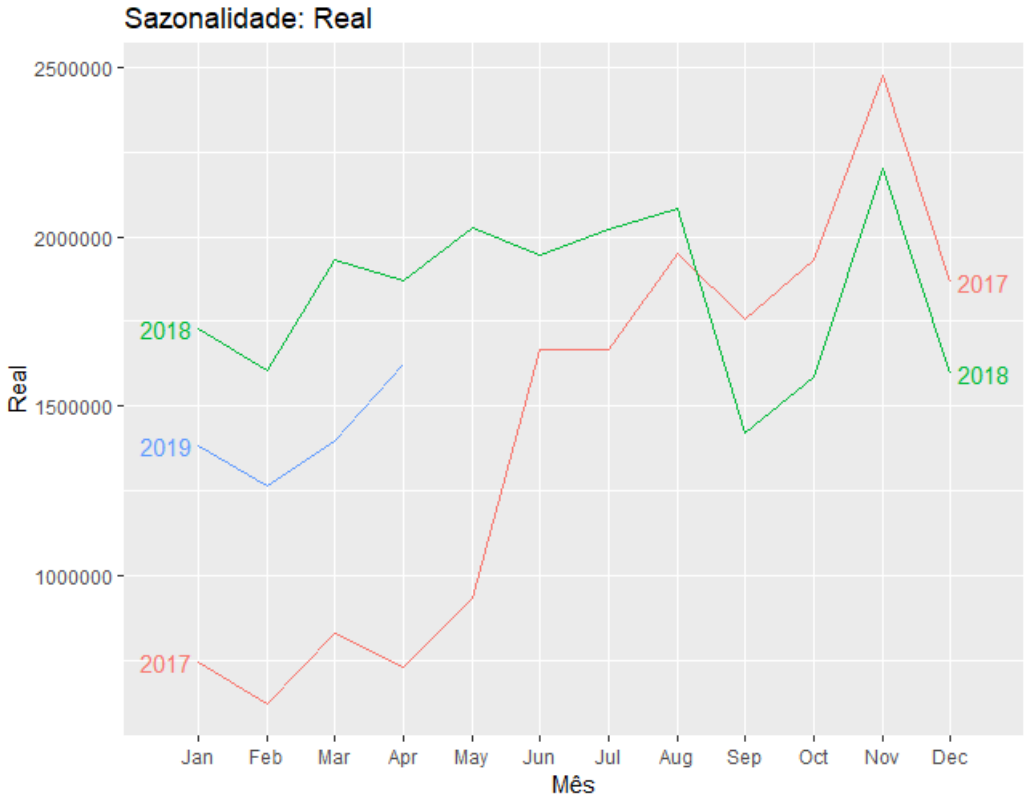

Figura 3.3: Sazonalidade mensal da série original ano a ano.

3.2 Ajustes dos modelos

Para o ajuste e avaliação dos modelos as observações foram separadas em treino e teste, sendo que após ser realizada a previsão e calculado o erro mais uma observação era colocada no conjunto de treino e os passos anteriores eram repetidos.

Os modelos ajustados foram:

- Regressões lineares, nas quais as variáveis explicativas eram:
  - previsão da plataforma (RL:1);
  - previsão da plataforma e sazonalidade (RL:2);
  - previsão da plataforma, sazonalidade e tendência (RL3).
- ARIMA(1,0,0) ou AR(1);
- Regressão dinâmica, na qual as variáveis explicativas eram a previsão da plataforma e a sazonalidade mensal. Os erros ϵ_t tinham como distribuição um ARIMA(0,1,0).

Para os três modelos lineares ajustados a suposição de que os resíduos não são autocorrelacionados foi violada (teste de *Ljung-Box*), dessa forma as previsões são ineficientes já que existem relações que não foram explicadas pelo modelo. Na figura 3.4 é ilustrada a situação dos modelos lineares. Os resíduos são referentes ao modelo linear que considera a previsão da plataforma e a sazonalidade (RL:2 na 3.1). Apesar da normalidade dos resíduos e da média dos mesmos ser zero, as autocorrelações das três primeiras defasagens foram significativas.

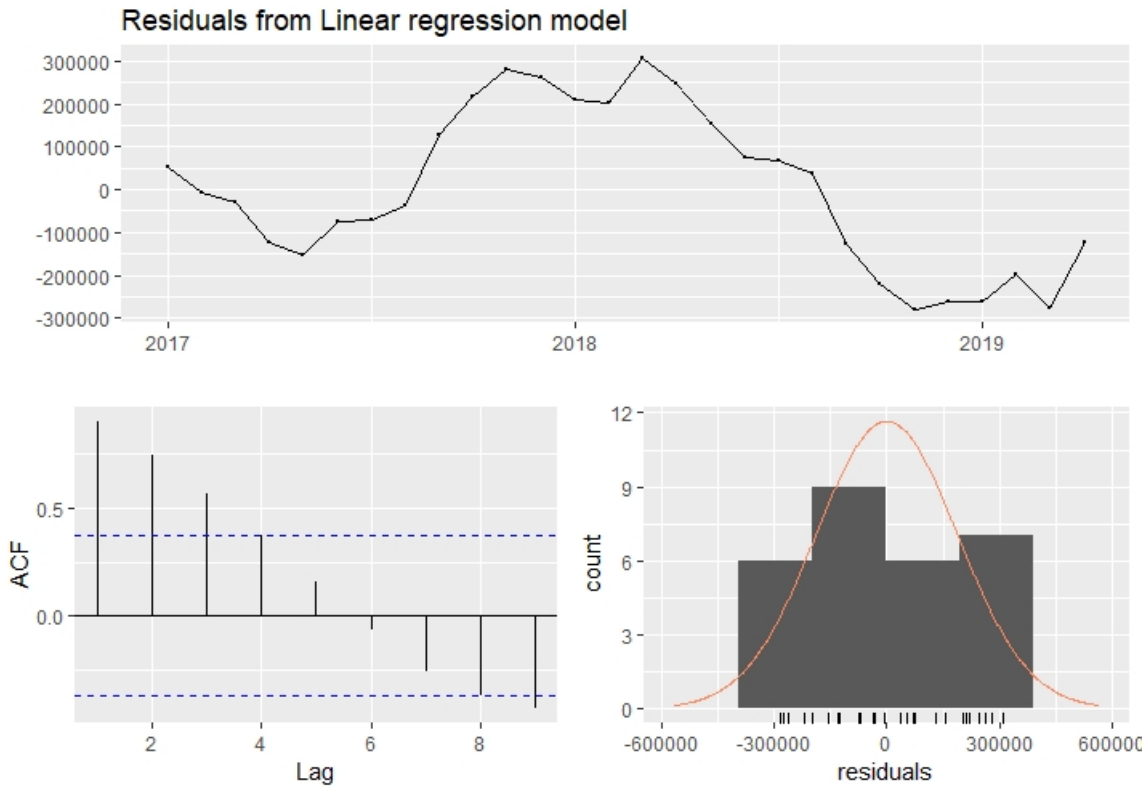

Figura 3.4: Análise dos resíduos do modelo RL:2.

O modelo ARIMA(1,0,0) ou AR(1) foi ajustado utilizando a função *auto.arima* do pacote *forecast*, essa função encontra os coeficientes que melhor se ajustam aos dados de acordo com o Critério de Informação de Akaike (AIC).

O modelo de regressão dinâmica também foi ajustado com a função *auto.arima*, mas dessa vez recebendo um argumento a mais, *newdata*, no qual são incluídos os regressores.

3.3 Seleção do modelo

A avaliação do modelo foi feita considerando a média dos erros relativos mensais (MAPE) de janeiro a abril de 2019.

Na tabela 3.1 estão os resultados obtidos dos modelos. O mês de janeiro teve erros altos em todos os modelos, uma possível explicação é a alteração que foi feita no *site* resultando em um tráfego muito maior do que o esperado.

Apesar da forte relação linear observada na análise descritiva, o modelo linear que considera somente a previsão inflada para prever o tráfego real obteve o pior desempenho. Quando componentes de séries temporais foram adicionadas os erros diminuíram em até três vezes.

O modelo autoregressivo obteve um bom resultado, mas adicionar a previsão inflada como variável explicativa e a sazonalidade mensal (modelo de regressão dinâmica) resultou em um menor

MAPE.

Dessa forma, o modelo escolhido para as previsões de tráfego orgânico desse site foi o modelo de regressão dinâmica com a estimativa da plataforma e a sazonalidade da séries real.

Modelo	**RL:1**	**RL:2**	**RL:3**	**Arima**	**Regressão dinâmica**
Janeiro	40,7	38,1	24,8	15,5	29,6
Fevereiro	52,4	28,7	9,7	26	6,2
Março	33,4	32,5	14,7	9	3,4
Abril	20,6	12,3	3,3	12,1	12,4
MAPE	36,8	27,9	13,1	16,2	12,9

Tabela 3.1: Tabela com o erro absoluto relativo mensal de cada modelo e os respectivos MAPE.

4 Próximos passos

Como próximos passos pretende-se:

- automatizar a análise, para que a cada mês as estimações sejam feitas sozinhas;
- aplicar essa análise para outros clientes e comparar resultados.

Referências Bibliográficas

[HYNDMAN e ATHANASOPOULOS, 2019] HYNDMAN, R.J. & ATHANASOPOULOS, G. *Forecasting: Principles and Practice*. OTexts, 2019. Disponível em: http://OTexts.com/fpp2. Acesso em: 10 de Maio de 2019.

[KANNAN e LI, 2017] KANNAN, P.K. & LI, H. A. *Digital marketing: A framework, review and research agenda.*. International Journal of Research in Marketing, 2017, Vol. 34, No. 1, p. 22-45.

[LEEFLANG et. al, 2014] LEEFLANG, P. S. H., VERHOEF, P. C., DAHLSTROM, P., FREUNDT, T. *Challenges and solutions for marketing in a digital era*. European Management Journal, 2014, Vol. 32, No. 1, p. 1-12.

[SCHOBER et. al, 2018] SCHOBER, P., BOER, C., SCHWARTE, L.A. *Correlation Coefficients: Appropriate Use and Interpretation*. Anesthesia & Analgesia, 2018, Vol. 126, No. 5.