**PONTIFÍCIA UNIVERSIDADE CATÓLICA DE MINAS GERAIS NÚCLEO DE EDUCAÇÃO A DISTÂNCIA**

**Pós-graduação *Lato Sensu* em Ciência de Dados e Big Data**

**Ivanilton Ferreira Bastos**

**Análise de series temporais e predição de vendas de veículos novos no Brasil utilizando técnica de Web Scrapping e ARIMA**

**Manaus**

**2021**

**Ivanilton Ferreira Bastos**

**ANÁLISE DE SERIES TEMPORAIS E PREDIÇÃO DE VENDAS DE VEÍCULOS NOVOS NO BRASIL UTILIZANDO TÉCNICA DE WEB SCRAPPING E ARIMA**

Trabalho de Conclusão de Curso apresentado ao Curso de Especialização em Ciência de Dados e Big Data como requisito parcial à obtenção do título de especialista.

**Manaus**

**2021**

**SUMÁRIO**

1. [Introdução 4](#_bookmark0)
   1. [Contextualização 4](#_bookmark1)
   2. [O problema proposto 5](#_bookmark2)
2. [Coleta de dados 5](#_bookmark3)
3. [Processamento/Tratamento dos dados 6](#_bookmark4)
4. [Análise e exploração dos dados 7](#_bookmark5)
   1. [Análise exploratória das vendas 7](#_bookmark6)
5. [Modelos preditivos com ARIMA](#_bookmark10) 10
6. Conclusão 14

# INTRODUÇÃO

# Contextualização

No final de 2008, tivemos o início de uma das piores crises econômicas desde a grande depressão. O governo brasileiro, em contrapartida, realizou vários programas de incentivos fiscais, no intuito de alavncar a economia do pais através da produção industrial e do comércio.

No período de 2008 a 2013 foi realizado a redução de IPI (imposto sobre produtos industrializados) que reduziu o valor de venda dos produtos industrializados, como eletrodomésticos e veículos novos, visando assim aquecer o comércio e a indústria do país.

Diante desse contexto, este trabalho visa utilizar técnicas de Web Scrapping e Modelos Preditivos em séries temporais dos dados obtidos do site HTTP://www.autoo.com. Para isso foi desenvolvido um script em Python utilizando diversas bibliotecas, tais como Pandas, Seaborn, Matiplotlib, Statsmodels e ARIMA, além das básicas de uso comum da linguagem.

# O problema proposto

O problema estudado consiste no uso de Web Scrapping, Análise Exploratória e Modelagem Preditiva para extração de informações importantes das séries temporais de vendas anuais e trimestrais, a fim de verificar os resultados obtidos pela redução do IPI e a sazonalidade entre os trimestres de vendas, bem como apontar os melhores períodos para as vendas de veículos novos no Brasil. Através dos dados obtidos, verificaremos se é possivel realizar predições de vendas para o Dataset.

Para isso, são analisadas as séries temporais de emplacamentos de veículos no período de 2008 a 2020. Assim, têm-se como objetivos dessa análise:

* + - Realizar a mineração dos dados para compor nosso Dataset;
    - Realizar a limpeza e ordenação dos dados;
    - Realizar a análise exploratória dos dados;
    - Separar os dados em trimestres;
    - Obter *insides* a partir do Dataset;
    - Criar modelos preditivos para a venda de veículos novos partir dos dados de emplacamentos anuais utilizando a biblioteca AUTO-ARIMA.

O período utilizado na análise exploratória e treinamento dos modelos de predição corresponde aos anos de 2008 a 2020. Para os dados de teste dos modelos treinados, foi utilizado 30% dos dados e o restante foi utilizado para o treinamento.

# Coleta de Dados

Para o desenvolvimento deste trabalho foi utilizado Jupyter Notebook e a biblioteca *BeautifulSoup* para a coleta de dados do site https://www.autoo.com. O site Autoo é uma das referências em estatísticas de vendas e emplacamentos de veículos no Brasil.

Para a coleta de dados foi criada a função buscar\_Tabela(), que se conecta ao site e, através de Web Scrapping, copia a tabela em HTML e cria o arquivo csv, Emplacamentos e retorna na tela a resposta sobre a requisição e criação do arquivo.

#Função para o Scraping de dados da Web

def buscar\_Tabela():

url = 'https://www.autoo.com.br/emplacamentos/'

print(url)

req = requests.get(url)

if req.status\_code == 200:

print('Requisição bem sucedida!')

content = req.content

soup = BeautifulSoup(content, 'html.parser')

table = soup.find(name='table')

table\_str = str(table)

tabela = 'Emplacamentos'

dfDados = pd.read\_html(table\_str)[0]

dfDados = dfDados.to\_csv(tabela + '.csv', encoding = 'utf-8')

print('Foi criado o arquivo de Dados ' + tabela+ '.csv')

A partir dos dados extraídos, foi criado a tabela dfEmplacamentos que possui a estrutura a seguir:

|  |  |  |
| --- | --- | --- |
| **Coluna** | **Descrição** | **Tipo** |
| Ano | Período para análise | Int64 |
| meses | Uma coluna para cada mês do ano | Float64 |
| Total de Vendas | Soma das vendas mensais | Float64 |

# Processamento/Tratamento de dados

Após a criação do Dataframe, foram realizados vários procedimentos para o tratamento e limpeza dos dados.

Foram verificados os tipos dos dados e os valores nulos foram substituidos pelo valor zero. A coluna criada pela biblioteca Pandas para indexar o Dataframe foi retirada para que a indexação fosse feita através da coluna “Ano”. A coluna “Total de vendas” foi excluída e recriada através da soma dos valores de vendas mensais, visando corrigir o tipo dos dados dessa coluna.

**3, Análise e exploração dos dados**

# Análise exploratória sobre as vendas mensais

Para a realização da análise exploratória dos dados, foram levantados os questionamentos a seguir:

1. Qual foi o melhor ano para as vendas de veículos no Brasil?

2. Qual foi o pior ano para as vendas de veículos no Brasil?

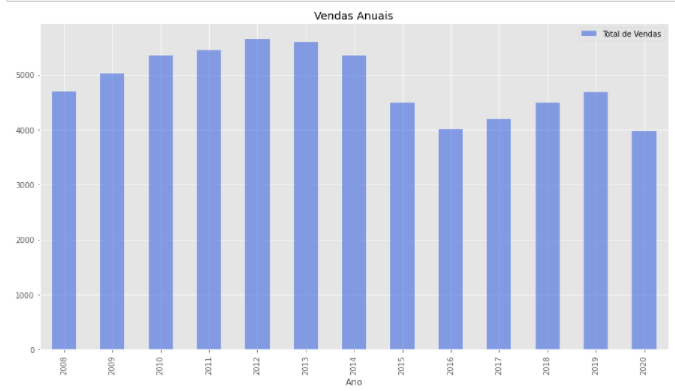
3. Como foi o comportamento das vendas de veículos por trimestre?

4. Quais os parâmetros estatísticos para as vendas anuais e trimestrais?

A fim de responder aos questionamentos, foram utilizadas as bibliotecas Matplotlib e Seaborn para plotados os gráficos abaixo:

A partir da análise da figura 1, podemos dizer que o melhor ano para as vendas de veículos novos no Brasil foi o ano de 2012, enquanto os piores anos foram 2020 seguido pelo ano de 2016.

Com a redução do IPI a partir do final de 2008, percebe-se uma tendência de alta que durou ate o final de 2012. No início de 2013, com o fim dos incentivos, vemos que as vendas de veículos retornam para uma tendência de baixa que dura até o final de 2016.



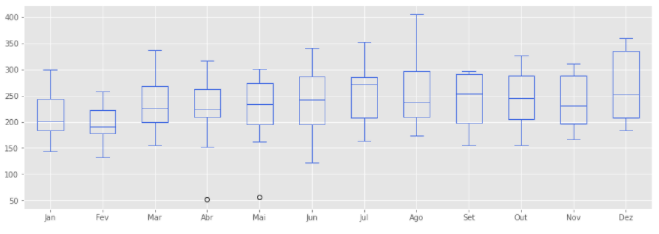
**Figura 1** – Gráfico em barras representando as vendas anuais.

Através do gráfico de caixas (Figura 2), podemos verificar um *outlier* no mês de maio. Esta discrepância é devido a falta de dados para o mês no ano corrente.

No mês de abril, verificamos a mínima histórica para as vendas, provavelmente puxada pelo mês no ano de 2020.

Já a máxima Histórica de vendas é verificada no mês de agosto, provavelmente influenciado pelos anos de 2012 e 2013.

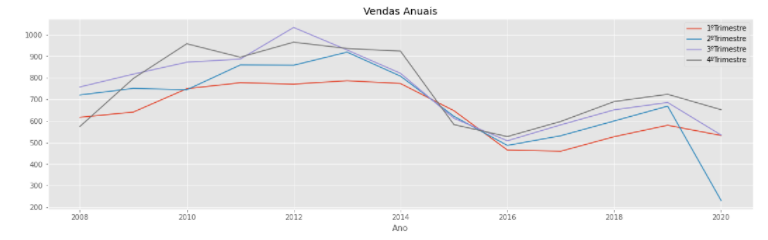
Podemos visualizar também, que as medianas de vendas se movem entre 200 e 250 mil vendas mensais. Observamos que o mês de dezembro possuí a maior dispersão de dados.



**Figura 2** – Bloxplot representando as vendas mensais.

Para a análise de tendência, os dados foram agrupados por trimestres conforme o grágico da figura 3:

A partir do segundo semestre de 2015, verifica-se o ínicio da tendência, a qual é interrompida no início do ano de 2020 a iminente crise gerada pela pandemia ocasionada pela COVid-19.

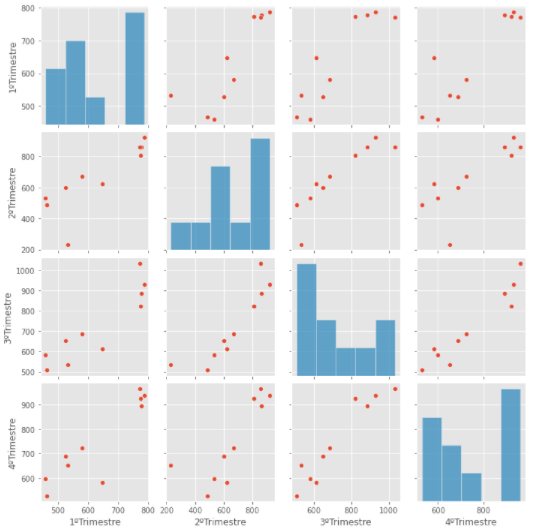


**Figura 3** – Gráfico de linhas por representando as vendas por trimestre.

Observa-se que os dois últimos trimestres de 2012 foram os responsáveis pelo recorde de vendas e que a diminuição das vendas de 2016 foi bem dividida ao longo do ano.

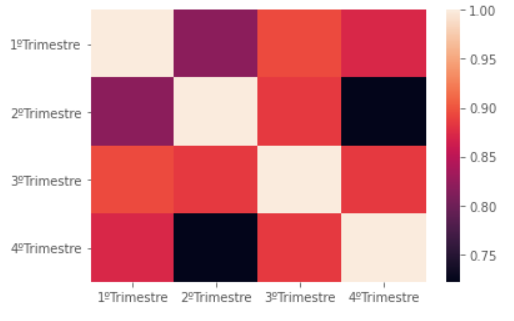
Podemos observar ainda que no ano de 2015 houve uma inversão nas vendas entre os trimestres.

O Pairplot (Figura 4) permite observar que a correlação entre os trimestres ganham força à medida que os trimestres passam, até chegar ao ápice de vendas no 4° trimestre.



**Figura 4** – Gráfico de linhas representando as vendas por trimestre.

Através do gráfico de correlação da figura 5, é visivel que existe uma correlação forte entre o final de cada semestre, mostrando que os segundo e quarto trimestres são onde ocorrem a maioria das vendas de veículos.



**Figura 5** – Gráfico de correlação representando as vendas por trimestre.

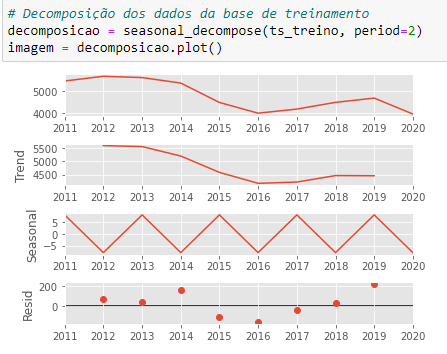
# Modelos Preditivos para a vendas de veículos anuais

Esta seção apresenta os modelos preditivos desenvolvidos em linguagem Python para as vendas anuais utilizando a biblioteca AUTO-ARIMA.

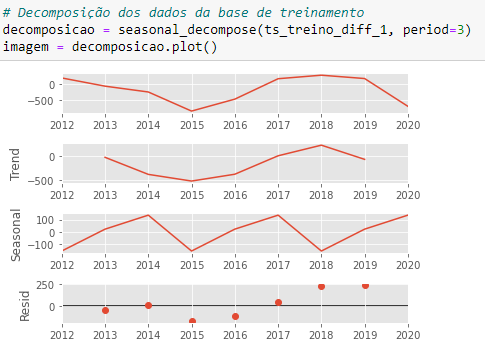
# Modelo Preditivo com AUTO-ARIMA

Para a utilização da biblioteca AUTO-ARIMA, foi realizado um breve estudo da série temporal, considerando aspectos tais como sazonalidade, estacionaridade e autocorrelação.

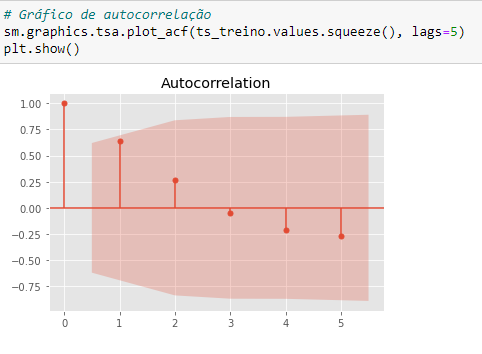
Decompondo-se a série temporal, obtém-se as componentes apresentadas na figuras 6 a 8.



**Figura 6** – Decomposição da série temporal para os dados de treino com período=2.

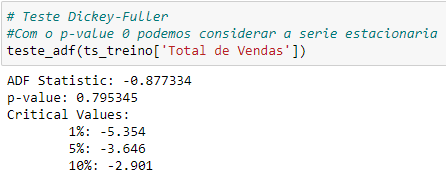


**Figura 7** – Decomposição da série temporal para os dados de treino com período=3.



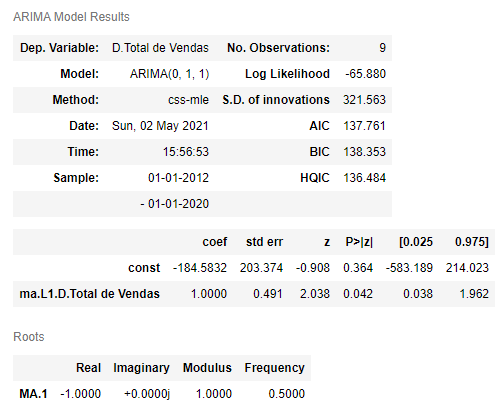
**Figura 8** – Autocorrelação dados de treino.

A estacionalidade foi verificada utilizando o Teste Dickey-Fuller, obtendo-se o seguinte resultado (Figura 9):

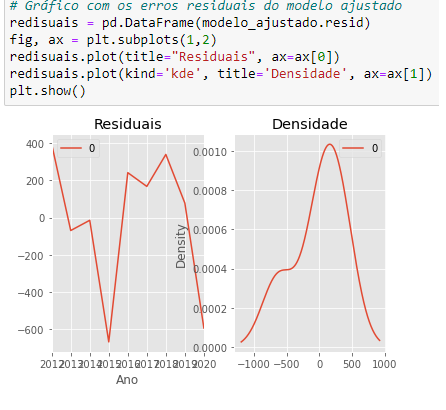


**Figura 9** – Teste Dicky-Fuller.

Para a implementação do modelo de médias móveis e regressão linear foi verificado a partir do teste de Dick-fuller que esta base de dados não é estacionária (Figura 10).



**Figura 10** – Resultado do modelo ARIMA.

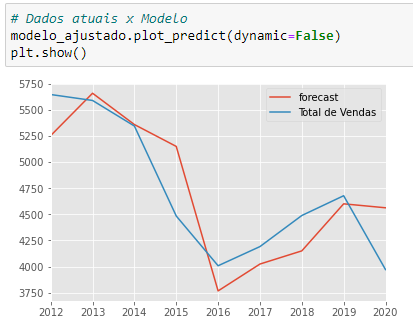
.

**Figura 11** – Residuais.

Observando o gráfico Residual (Figura 11) verificamos que existe um grande densidade nos dados residuais, o que dificultará a previsão correta dos dados.

A Figura 12 apresenta o resultado do nosso modelo. A previsão mostra algumas disprepancias em realação ao total de vendas, porém não é observado o overfiting em nosso modelo de previsão.

A queda acentuada em 2020 é relativa ao início da Pandemia, podemos observar na previsão uma correção das vendas em relação a este período no ano de 2020.



**Figura 12** – Resultado do modelo gerado pelo ARIMA.

# Conclusões

Através da técnica de *Scrapping* foi obtida a série temporal para o emplacamento mensal dos últimos anos.O emplacamento mensal reflete as vendas mensais de veículos novos no Brasil.

Após as análises realizadas, conclui-se que o melhor ano para as vendas de veículos novos no Brasil foi o de 2012, enquanto que o ano de 2016 apresentou o menor índice de vendas.

Observando o gráfico trimestral, pode-se perceber que os melhores trimestres para a venda de veículos são os 4º trimestres. sendo os 1º trimestres os que apresentam menor índice de vendas. Foi observado também que em 2015 houve uma inversão de vendas ao longo dos trimestres.

Ao iniciar a análise da série temporal, constata-se que, para uma análise mais precisa, é necessário a utilização de Dataset mais extenso tanto para o treinamento do modelo e decomposição sazonal quanto para a predição dos dados.

A decomposição sazonal demonstrou que, a partir de 2015, as vendas de veículos novos estão em uma tendência de alta e que, no período análisado, foi observado alta sazonalidade entre os anos.

O mercado de vendas de veículos novos no Brasil já sofria com a queda no período de 2014 a 2016. A partir de 2017, o mercado se preparava para um período de alta nas vendas, porém foi interrompido com a crise gerada pela pandemia da COVid-19, marcando assim um novo período de baixa a partir de 2020.

Para a predição foram testados os modelos de Regressão linear, Random Forest e o ARMA (ARIMA): os dois primeiros apresentaram altos níveis de erros, provavelmente pela baixa quantidade de dados do nosso DataSet; o modelo ARIMA foi escolhido por possibilitar a manipulação de configurações de forma mais prática para o usuário.

Links: [IvaniltonBastos/TCC-PUCMINAS (github.com)](https://github.com/IvaniltonBastos/TCC-PUCMINAS/tree/main)

<https://youtu.be/NDMpXVqtBqE>