Link para o arquivo no Colab: <https://colab.research.google.com/drive/1vyMhMgspLq8QFDttjp2F7wa1QKOrMF7w?usp=sharing>

Lá, toda a explicação do passo a passo está detalhada. Copiei abaixo para consulta rápida:

**ABERTURA**

Para este projeto, viso responder às perguntas:

* Consumidores confiantes sobre o mercado compram mais carros?
* Dado a quantidade de carros produzidos e vendidos este ano, assim como o índice de confiança do consumidor, quantos carros serão vendidos no próximo ano?

Para isso, olharemos para três variáveis: A quantidade de automóveis produzidos em determinado ano, quantidade de vendas de automóveis/ano e o índice de confiança do consumidor (ICC).

**DATA ACQUISITION**  
Todos os dados foram retirados da Instituto de Pesquisa Econômica Aplicada (Ipea), acessável pelo link: http://www.ipeadata.gov.br/Default.aspx

Os CSVs baixados foram os seguintes:

Vendas de autoveículos no mercado interno

Produção - automóveis - montados - qde.

Índice de confiança do consumidor

Todos foram baixados manualmente. Infelizmente, tive dificuldades em juntar eles por Python, então fiz a junção manualmente, já que todos seguem a mesma estrutura de tempo. Até a entrega do projeto final, irei refazer esta etapa por código.

O arquivo gerado, que será base de todo o código à seguir, está disponível aqui: https://drive.google.com/file/d/1T4C8eWlPsth6u4IlXtaqz1V\_583s33VO/view?usp=sharing

**DATA WRANGLING**  
  
*~ código ~*

Com os dados limpos, seguimos para o projeto.

Dicionário de termos:

**1. \*\*Vendas de autoveículos no mercado interno:\*\*** Total de vendas de carros no mercado brasileiro por mês. Dados extraídos da Associação Nacional dos Fabricantes de Veículos Automotores, Anuário Estatístico da Indústria Automobilística Brasileira ou Carta Mensal (Anfavea);

**2. \*\*Produção - automóveis - montados - qde.:\***\* Produção de carros no Brasil. Obs.: "Refere-se apenas a carros de passeio / passageiros e de uso misto, não englobando veículos comerciais leves (caminhonetes de uso misto, utilitários e caminhonetes de carga) nem veículos comerciais pesados", também extraído da Anfavea;

**3. \*\*Índice de confiança do consumidor\*\*:** "O Índice de confiança do consumidor (ICC) tem como objetivo identificar o sentimento dos consumidores levando em conta suas condições econômicas atuais e a expectativa em relação à sua situação econômica futura. O índice é calculado a partir de um questionário e, em cada pergunta, são computadas as diferenças entre as frequências de respostas positivas e negativas, sendo o índice final a média simples dessas diferenças. Variando de 0 (pessimismo total) a 200 (otimismo total), ele é composto a partir de dados segmentados por nível de renda, sexo e idade. O ICC é composto por outros dois índices: o índice de condições econômicas atuais (ICEA) e o índice de expectativas do consumidor (IEC)". Calculado pela Federação do Comércio do Estado de São Paulo, Pesquisa Conjuntural do Comércio Varejista da Região Metropolitana de São Paulo (Fecomercio SP).

Quantidade de observações (cada uma sendo um mês): 365

**ANÁLISE EXPLORATÓRIA**  
  
Começamos por identificar as correlações. Aqui, conseguimos responder à primeira pergunta de negócio.

*~ código ~*

Matriz de Correlação:

Producao Vendas ICC

Producao 1.000000 0.892396 0.659588

Vendas 0.892396 1.000000 0.641574

ICC 0.659588 0.641574 1.000000

Como observado, todas as combinações geram resultados satisfatórios para nossa análise. A venda está fortemente ligada à produção, indicando que as fabricantes de carro fazem pesquisas para produzir a quantidade ideal por ano, evitando excessos. Neste, a correlação é de 0.89.

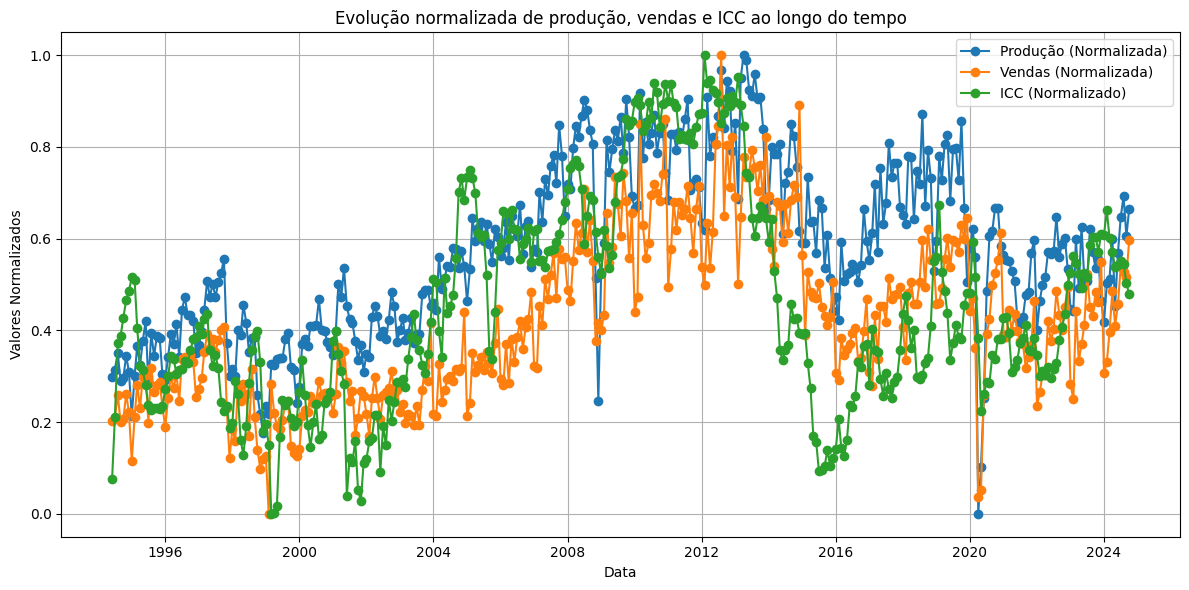
Também, identificamos que uma parcela importante da motivação dos consumidores é a confiança no mercado. Os consumidores precisam ter indicações que a economia estará estável no futuro antes de comprar um bem deste valor. Neste, a correlação é de 0.64.

Isto responde a primeira pergunta: consumidores confiantes no mercado têm mais chances de comprar um carro novo.

**VISUALIZAÇÃO DE DADOS**

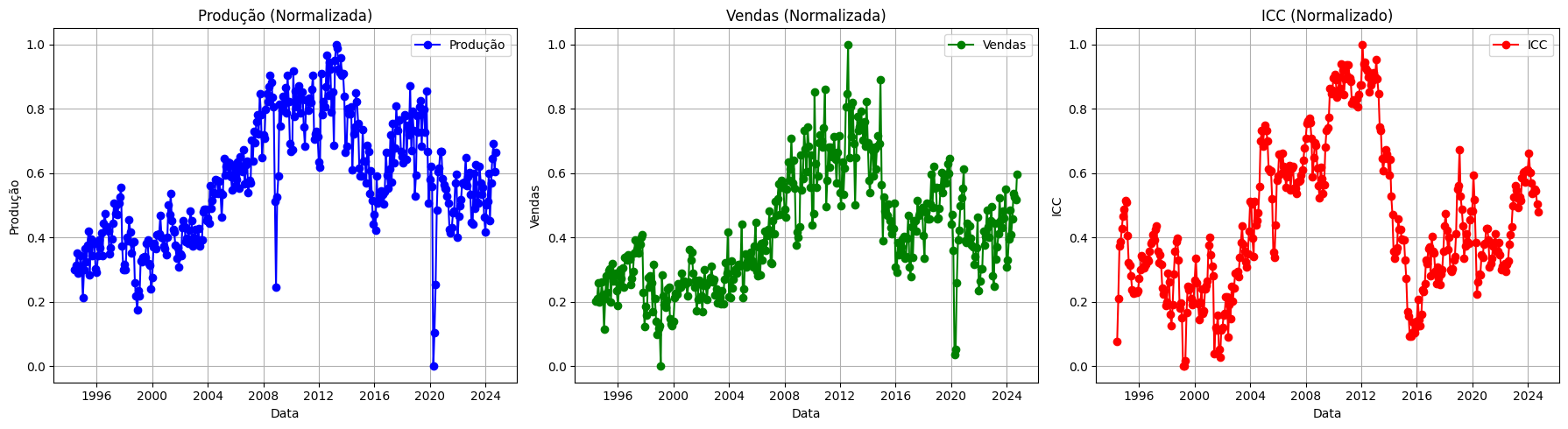
Vejamos em um gráfico o histórico dessas variáveis, corroborando a matriz de correlação:

*~ código ~*

**

Claramente, as variáveis percorrem um caminho bastante similar. Porém, pela quantidade de observações e linhas, é difícil seguir uma delas como referência. Por isso, abaixo plotamos as três de forma separada:

*~ código ~*



Finalmente, chegamos a conclusão de que os dados estão em condições de serem usados em um modelo preditivo.

**BORUTA E AVALIAÇÃO DE MODELOS**

*~ código ~*

Aceitadas:

Vendas

Rejeitadas:

Producao

ICC

O BORUTA recomendou rejeitar duas das três variáveis. Como esse projeto é baseado em usar uma variável para tentar prever a outra, continuei fazendo testes para garantir que essa seria a melhor ideia.

Como a variável Produção está muito relacionada com a Venda, e geralmente a venda influencia a produção (e não o contrário), experimentei por deixar ela de lado e ver pelo Random Forest Regressor se isso aumenta a acurácia do modelo.

*~ código ~*

Desempenho com todas as variáveis:

- Mean Squared Error (MSE): 6284470.0979

- R² Score: 0.9978

Desempenho sem Producao:

- Mean Squared Error (MSE): 3333224.6039

- R² Score: 0.9988

A remoção da variável produção impacta positivamente o modelo. Como podemos ver, o Erro Quadrático Médio (MSE em inglês) reduz quando retiramos ele. Também, R² melhorou levemente.

Por isso, iremos usar essas duas variáveis daqui para frente.

*~ código ~*

Resultados - Mean Squared Error (MSE):

Random Forest: 3333224.6039

Regressão Linear: 0.0000

Gradient Boosting: 2172813.3933

Resultados - R² Score:

Random Forest: 0.9988

Regressão Linear: 1.0000

Gradient Boosting: 0.9992

*~ código ~*

Validação cruzada:

Random Forest:

- R² Médio: 0.9975

- R² Desvio Padrão: 0.0033

Regressão Linear:

- R² Médio: 1.0000

- R² Desvio Padrão: 0.0000

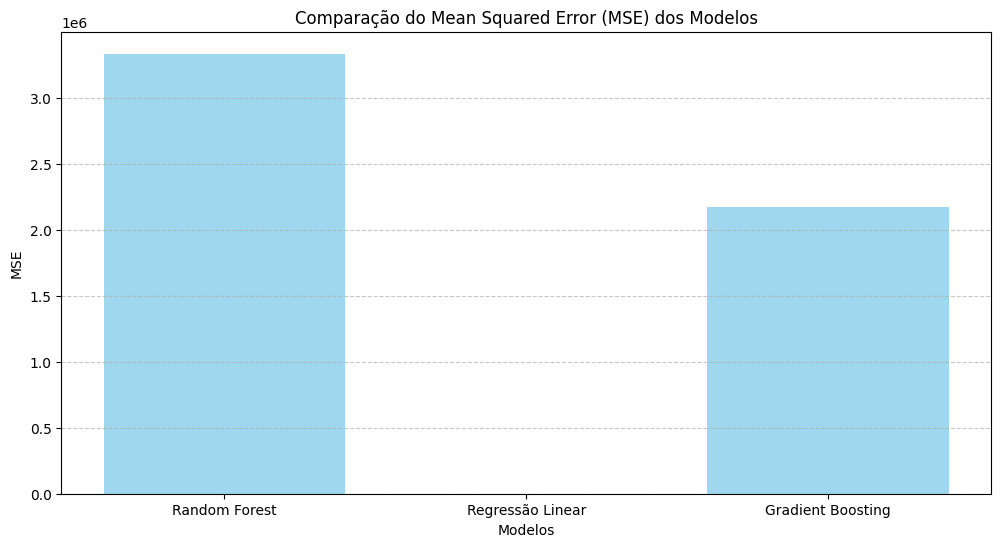
Gradient Boosting:

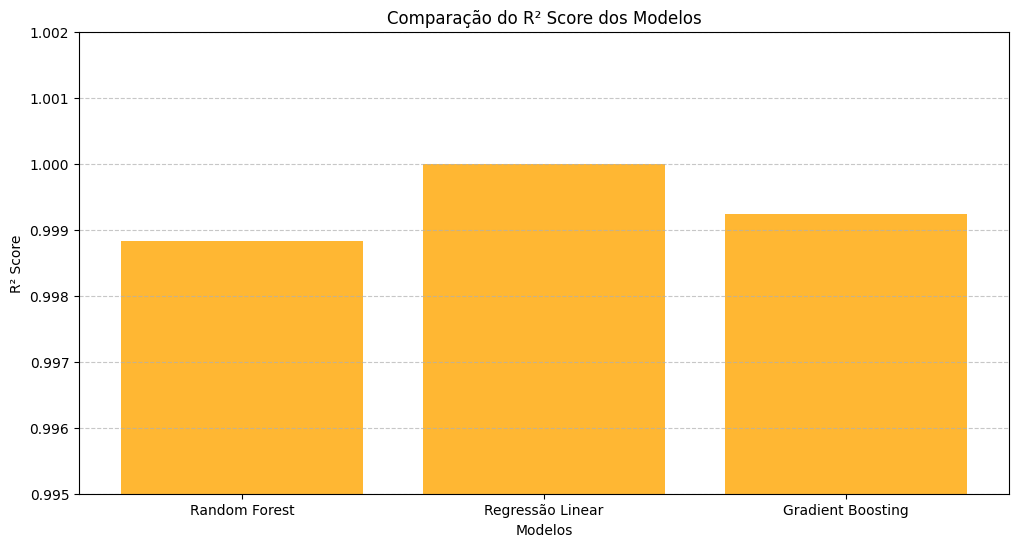
- R² Médio: 0.9981

- R² Desvio Padrão: 0.0028

**GRÁFICO COMPARATIVO DO DESEMPENHO DOS MODELOS**

*~ código ~*





A \*regressão linear\* me deu um resultado estranho, já que na matriz de correlação, essas duas variáveis estavam em 0.64. Depois de pesquisar muito e da ajuda de alguma inteligência artificial famosa, me foi sugerido fazer alguns testes, incluindo o abaixo:

*~ código ~*

Matriz de Correlação:

Vendas ICC

Vendas 1.000000 0.641574

ICC 0.641574 1.000000

VIF (Variance Inflation Factor):

Variável VIF

0 Vendas 17.261798

1 ICC 17.261798

Como o VIF estpa muito elevado, algo está impedindo o modelo de entender a contribuição de cada variável. Por isso, deixaremos ele de lado para usar uma das variáveis restantes.

Decidi então por usar o Gradient Boosting, já que ele me traz uma acurácia maior. Acurácia é o que preciso para uma previsão de um valor futuro.

**CONCLUSÃO E AVALIAÇÃO DAS MÉTRICAS**

Como percebemos pelos gráficos e cálculos, o Gradient Boosting segue como o melhor modelo atualmente. A métrica principal usada para tomar essa conclusão é o Erro Quadrático Médio, já que quero fazer uma previsão e isso necessita de acurácia.