UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS JARDINS DE ANITA

CIÊNCIA DE DADOS

ERICK COUTINHO

RELATÓRIO: ANÁLISE DE REGRESSÃO I

ITAPAJÉ, CE

2023

1. INTRODUÇÃO

1.1 DATASET

O dataset estudado e utilizado nesse relatório foi o: “Conjunto de Dados de Saúde e Demografia Global”, representando uma ferramenta inestimável para pesquisadores que buscam compreender a complexa interação entre fatores demográficos e saúde pública em escala global. Cada coluna deste conjunto de dados conta a história única de nações, fornecendo uma visão detalhada das variáveis que influenciam a esperança de vida e a saúde da população. Esta riqueza de informações oferece a oportunidade de conduzir pesquisas e projetos, contribuindo para a formulação de políticas baseadas em evidências e para a resolução de desafios significativos em saúde pública e epidemiologia, em um mundo em constante mudança.

Conhecendo o Dataset -

- **País:** Nome dos países.
- **Ano:** Respectivo ano das informações.
- **Status:** Status de desenvolvimento, seja “Desenvolvido” ou “Em Desenvolvimento”, que molda o curso da saúde.
- **Expectativa de vida:** Expectativa de vida no país em determinado ano.
- **Mortalidade de adultos: P**robabilidades de sobrevivência entre 15 e 60 anos de idade por 1.000 habitantes.
- **Mortes Infantis:** Número de mortes infantis por 1.000 nascidos vivos.
- **Álcool:** Consumo médio de álcool em litros per capita.
- **Percentagem de Despesas:** Despesas de saúde como uma percentagem do PIB de um país.
- **Hepatite B:** Cobertura vacinal para a Hepatite B.
- **Sarampo:** Impacto desta doença com o número de casos notificados por 1.000 habitantes.
- **IMC:** Índice de Massa Corporal médio.
- **Mortes de menores de cinco anos:** Mortalidade infantil com o número de mortes de menores de cinco anos por 1.000 nascidos vivos.
- **Poliomielite:** Cobertura vacinal contra a poliomielite.
- **Despesas Totais:** Despesas totais com saúde como uma percentagem do PIB.
- **Difteria:** Cobertura vacinal contra a difteria.
- **VIH/SIDA:** Prevalência do VIH/SIDA como percentagem da população.
- **PIB:** Produto Interno Bruto.
- **População:** Fluxo e refluxo da população de uma nação.
- **Magreza de 1 a 19 anos:** Magreza entre crianças e adolescentes de 1 a 19 anos.
- **Magreza de 5 a 9 anos:** Magreza entre crianças de 5 a 9 anos.
- **Composição dos Recursos do Rendimento:** Índice composto que reflete a distribuição do rendimento e o acesso aos recursos.
- **Escolaridade:** Média de anos de escolaridade.

2. DESCRIÇÃO DAS ATIVIDADES

O trabalho engloba várias etapas cruciais no processo de análise de dados e modelagem estatística. De acordo com as instruções, inicialmente, são realizados pré-processamentos de dados, nos quais os dados são preparados de acordo com o tipo de variável, incluindo a limpeza e a formatação necessárias. Em seguida, uma análise exploratória dos dados é conduzida para identificar relações entre as variáveis e captar insights iniciais, criação de gráficos e tabelas para apresentar as variáveis e demonstrar suas propriedades e singularidades.

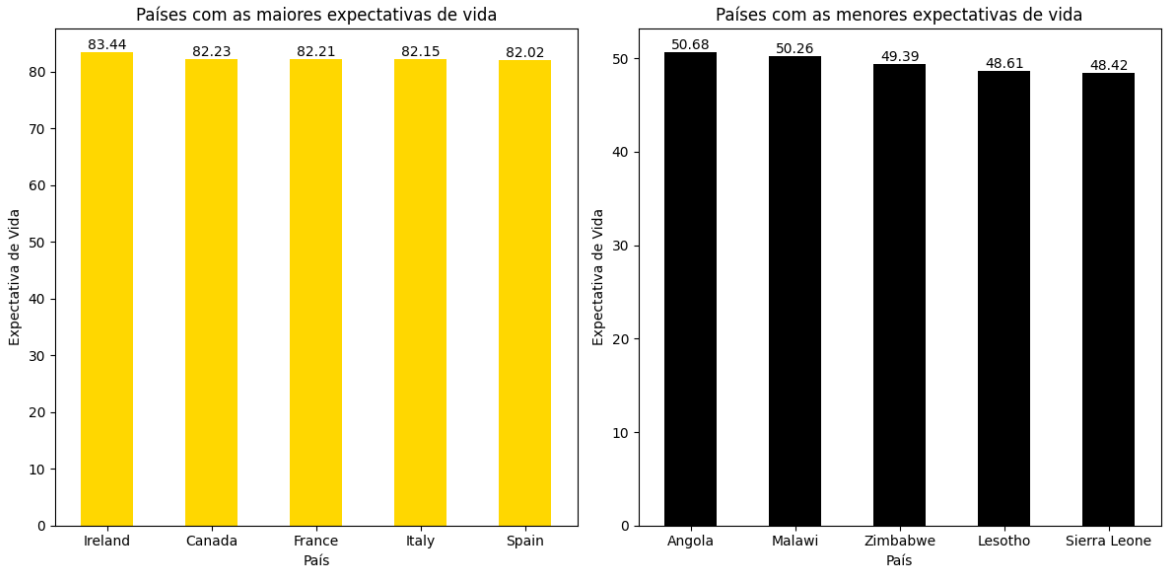
Posteriormente, ocorre o ajuste do modelo estatístico apropriado, que é crucial para a interpretação das estimativas e medidas de qualidade do ajuste. Além disso, são realizadas análises de variância (ANOVA) para avaliar diferenças significativas entre grupos ou categorias. A verificação de multicolinearidade, análise de resíduos e a aplicação de testes são realizadas para verificar as suposições do modelo, garantindo sua robustez.

Conforme necessário, etapas de seleção e transformação de variáveis são implementadas para otimizar a qualidade do modelo estatístico. Essas atividades são fundamentais para garantir a validade e a confiabilidade dos resultados obtidos na análise estatística, proporcionando uma base sólida para conclusões e recomendações subsequentes.

3. RESULTADOS

3.1 ANÁLISE EXPLORATORIA

- Inicialmente vamos analisar as maiores e menores expectativas de vida dentre os países do dataset, para conhecermos melhor a variável que usaremos como variável resposta.



No gráfico observado acima, temos lado a lado os 5 países com as maiores e menores expectativas de vida, respectivamente. Analisando o gráfico, pode-se perceber que as maiores expectativa de vida são de países europeus, sendo o que possui a maior expectativa o país: Irlanda, enquanto os 5 com as menores expectativas de vida, todos são estão localizados no continente africano, sendo o menor: Serra Leoa, notando-se a grande diferença entre a qualidade de vida nos diferentes continentes, como já era esperado.

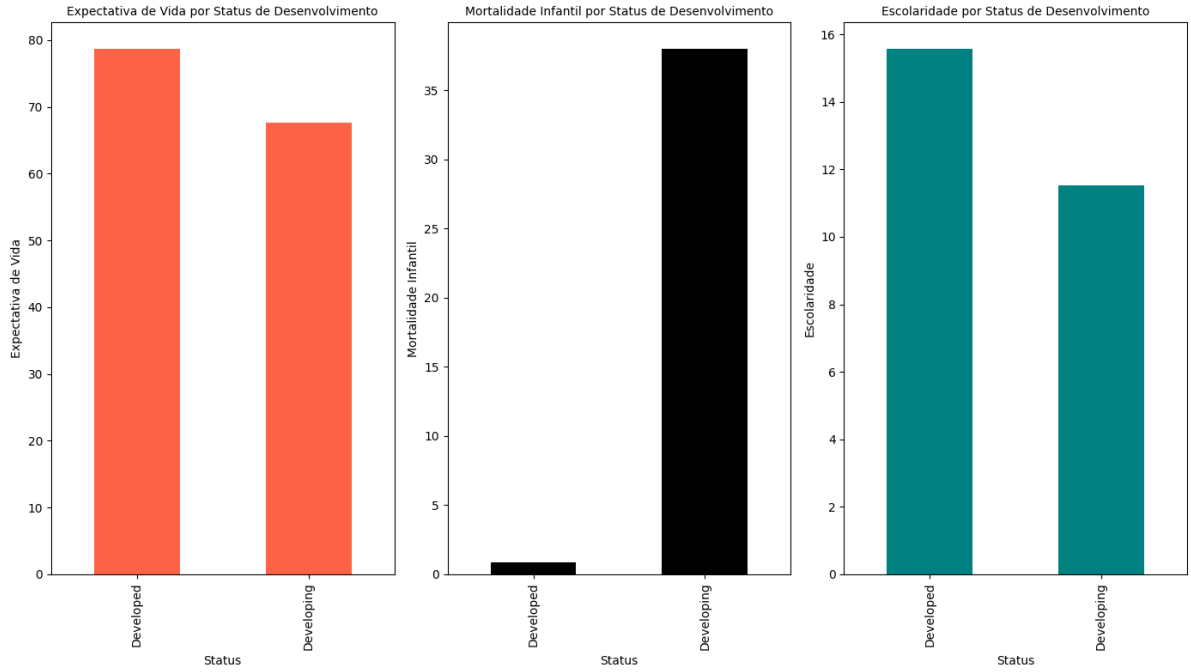
- A segunda análise realizada, foi a expectativa de vida do Brasil e da Argentina com o passar dos anos, para efeito de comparação, com dois países da América do sul.

Comparando esses dois países durante esse período, podemos observar que a expectativa de vida é relativamente semelhante, com pequenas variações ao longo dos anos. Ambos os

países experimentaram um aumento na expectativa de vida de 2002 a 2014. O Brasil começou com uma expectativa de vida ligeiramente menor em 2002, mas ambos os países alcançaram valores bastante próximos em 2014.

Essa análise nos mostra que, durante esse período, tanto o Brasil quanto a Argentina tiveram melhorias na expectativa de vida de sua população.

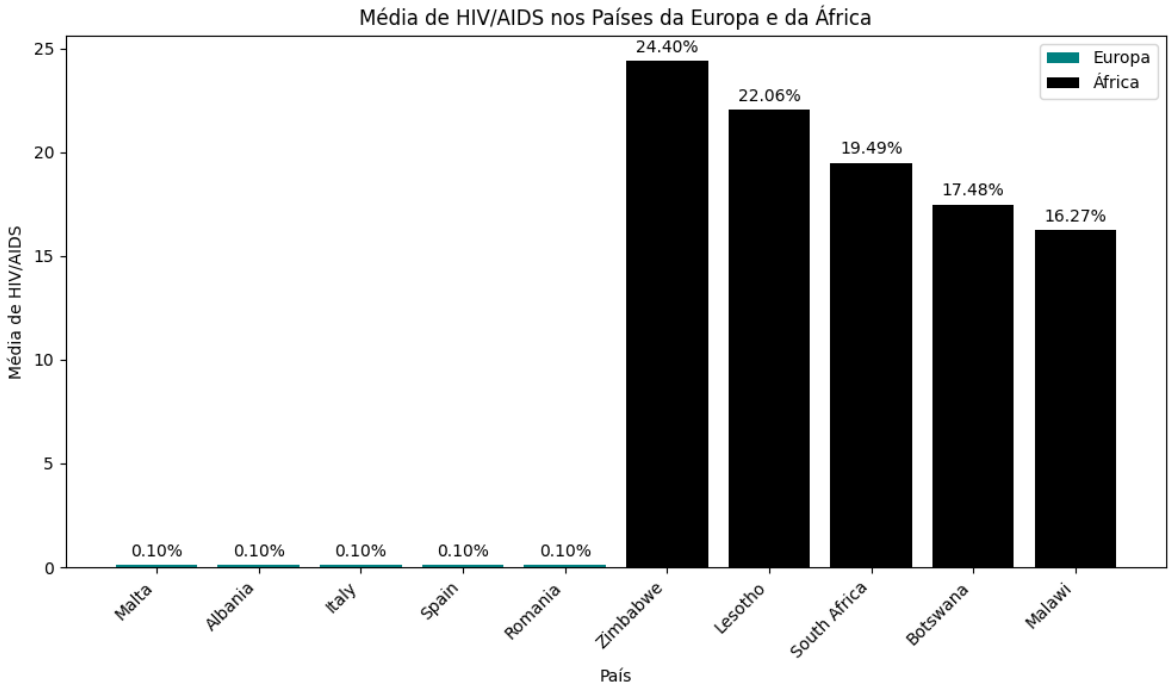
- Comparando grupos: países desenvolvidos e países em desenvolvimento -



No gráfico analisado, temos comparações de estatísticas sobre mortalidade, Expectativa de vida e escolaridade de países, separados em grupos: desenvolvidos e em desenvolvimento.

Percebemos a grande diferença dessas estatísticas de um grupo para outro, em todos as estatísticas, mas principalmente na “Mortalidade Infantil”.

- A última análise da parte exploratória, foi sobre a porcentagem de HIV na população do país, separando os 5 países de cada continente (europeu e africano) com a maior porcentagem da população contendo HIV.



Como podemos observar, a prevalência do HIV/AIDS é consideravelmente mais alta nos países africanos listados, com Zimbabwe e Lesoto liderando a lista. Por outro lado, os países europeus listados têm médias muito baixas, todas elas com uma média de 0.1. Essa diferença nas médias de HIV/AIDS reflete as disparidades na prevalência da doença entre as regiões da África e da Europa.

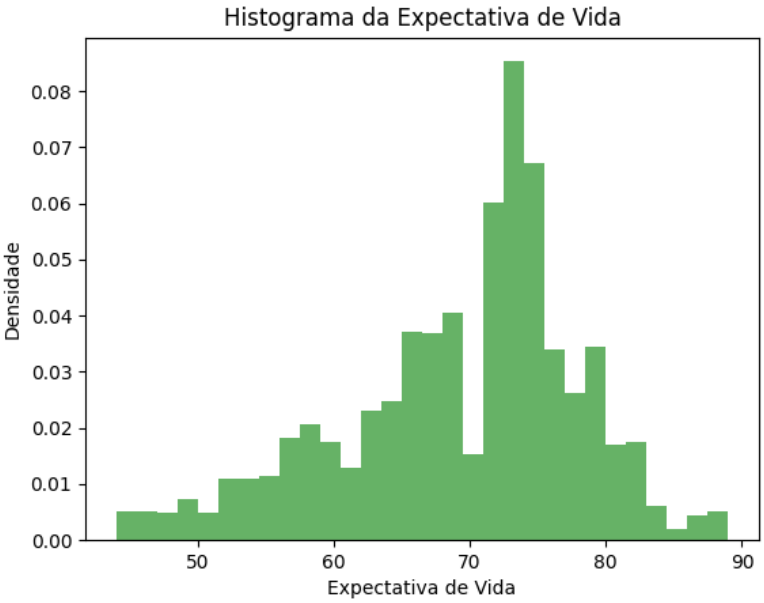
3.2 AJUSTE DO MODELO

Neste tópico, foi realizada uma análise de regressão linear múltipla com o objetivo de prever a expectativa de vida com base em seis variáveis independentes: mortalidade de adultos, mortes infantis, magreza em crianças e adolescentes, HIV/AIDS, Produto Interno Bruto (GDP) e consumo de álcool (Alcohol). A análise incluiu a verificação da normalidade da variável de resposta, o treinamento e ajuste do modelo, e a avaliação da qualidade do ajuste usando métricas como R^2, MSE e MAE. A análise é concluída com um gráfico de dispersão que compara os valores reais e previstos da expectativa de vida.

- Ajustando o modelo:

Para realizar uma análise de regressão linear, a variável reposta (y) não precisa obrigatoriamente ter uma distribuição normal. No entanto, é importante que os erros do modelo (ou seja, as diferenças entre as previsões e os valores reais) tenham uma distribuição aproximadamente normal. Isso ajuda a garantir que o modelo funcione bem.

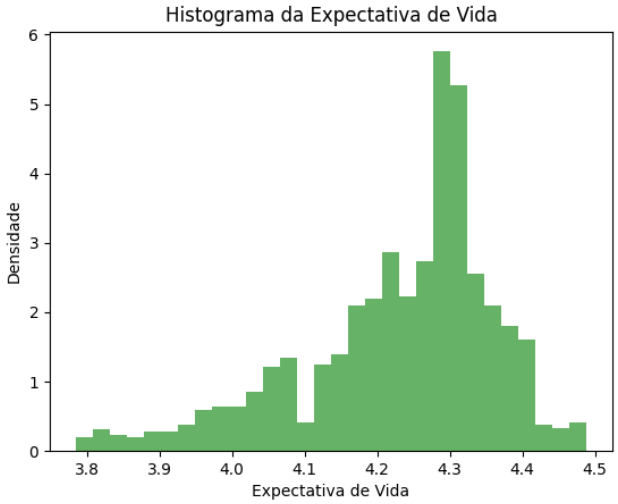
- Distribuição da variável resposta:



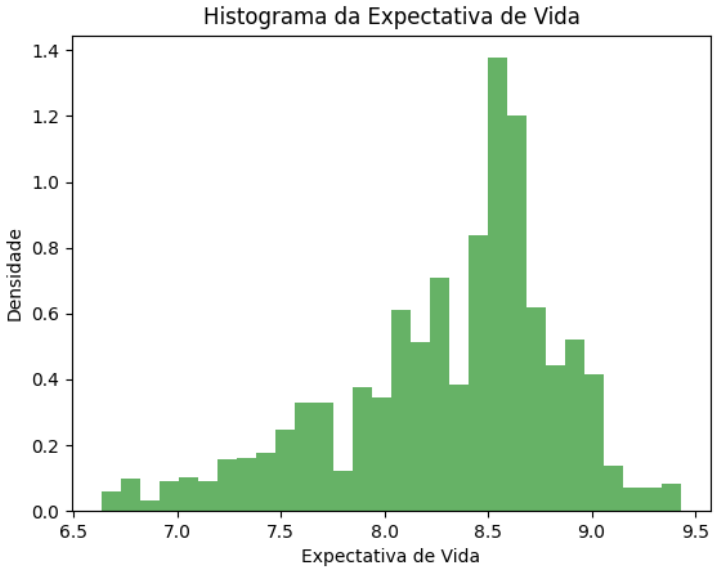
Observando o histograma, percebemos que a distribuição da variável resposta(y) não segue uma distribuição normal. Em casos de não normalidade, você pode considerar abordagens alternativas, como a transformação dos dados ou o uso de modelos estatísticos mais robustos a desvios da normalidade, dependendo da natureza dos seus dados e da pergunta de pesquisa.

- Transformando o Y:

Distribuição usando transformação Logarítmica:



Distribuição usando transformação de raiz quadrada:

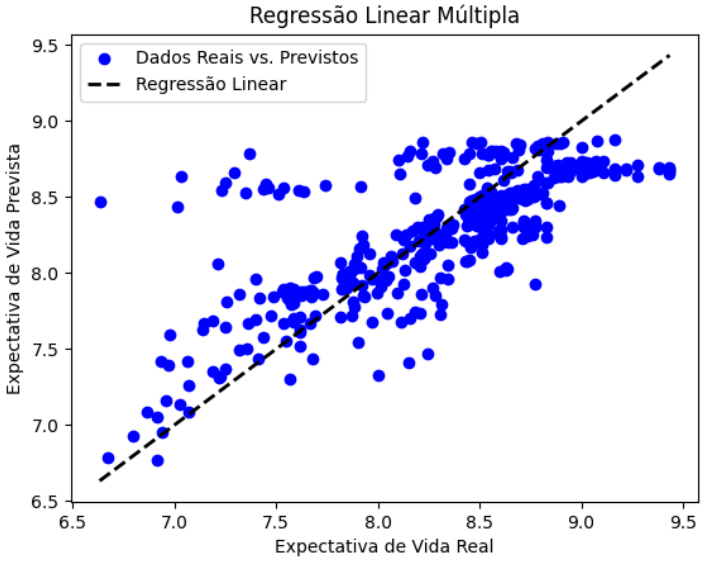


Mesmo após as transformações a variável resposta (y) não se aproximou de uma distribuição normal, isso pode indicar desafios na adequação de um modelo de regressão linear aos dados. Nesse cenário, considerar outras abordagens estatísticas, como modelos não lineares, modelos robustos ou técnicas mais avançadas, pode ser necessário.

Vamos continuar o ajuste ao modelo com y sem transformações. Lembrando que é importante adaptar a análise estatística à natureza dos dados e às necessidades da sua pesquisa.

- Gráfico de dispersão e avaliação do modelo:

Comparação os valores reais e previstos da expectativa de vida:



- Avaliação do modelo

R^2:	0.7011219630496898
MSE:	23.610939408965702
MAE:	3.631772911415011
Coeficientes:	[-2.79479139e-02 -3.20709000e-03 -3.12593146e-01 -4.46815747e-01 1.34524823e-04 3.65205818e-01]
Intercept:	73.9484729921444

- R^2 (R-squared): 0.7011, o que significa que aproximadamente 70,11% da variabilidade na variável dependente (Expectativa de Vida) é explicada pelas variáveis independentes incluídas no modelo. Um R^2 de 1 indicaria um ajuste perfeito.
- MSE (Mean Squared Error): 23.6109, que é a média dos quadrados das diferenças entre os valores reais e os valores previstos pelo modelo. Quanto menor o MSE, melhor o modelo ajusta-se aos dados.
- MAE (Mean Absolute Error): 3.6318, que é a média dos valores absolutos das diferenças entre os valores reais e os valores previstos pelo modelo. O MAE fornece uma ideia da magnitude dos erros de previsão. Quanto menor o MAE, menor é a magnitude dos erros.
- Coeficientes: Os coeficientes das variáveis independentes são os seguintes:

1. Mortalidade de Adultos: -0.0279
2. Mortes Infantis: -0.0032
3. Magreza em 1-19 anos: -0.3126
4. HIV/AIDS: -0.4468
5. GDP (Produto Interno Bruto): 0.0001
6. Alcohol (Consumo de Álcool): 0.3652

Esses coeficientes representam o impacto de cada variável independente na expectativa de vida. Por exemplo, um aumento na mortalidade de adultos está associado a uma diminuição de aproximadamente 0.0279 na expectativa de vida, mantendo as outras variáveis constantes.

O intercepto é de cerca de 73.9485 anos e representa a expectativa de vida quando todas as variáveis independentes são iguais a zero.

3.3 ANOVA

A ANOVA (Análise de Variância) é uma técnica estatística usada para comparar as médias de três ou mais grupos diferentes para determinar se há diferenças significativas entre eles. É uma ferramenta poderosa para a análise de dados e é comumente usada em diversas áreas, como ciências sociais, biologia, economia e engenharia. A ANOVA é uma extensão do teste t, que é usado para comparar duas médias.

Realizamos uma análise de variância (ANOVA) com base na expectativa de vida separada em diferentes anos, verificando se os diferentes anos têm algum efeito significativo na expectativa de vida dos países.

- Resultados

Não há diferenças significativas entre os grupos (p >= 0.05)

Estatística F:	0.9666411151758963
Valor P:	0.48553734016022576

"Não há diferenças significativas entre os grupos (p >= 0.05)": Este é um resultado importante da ANOVA. Significa que, com base na análise estatística, não há evidências suficientes para concluir que existem diferenças significativas na expectativa de vida entre os anos.

A estatística F é um valor calculado durante a ANOVA e é usada para determinar se existe alguma diferença significativa entre os grupos. O valor de Estatística F é próximo a 1, o que geralmente indica que não há diferenças significativas entre os grupos. O valor P (0.4855) é o resultado do teste de significância. Um valor P maior que 0,05 (5%) é geralmente considerado como evidência de que não há diferença significativa entre os grupos, sendo esse o caso.

Para atender os pressupostos da ANOVA (Normalidade e Homogeneidade) foram feitos o Teste de Levene e o teste de Shapiro-Wilk:

- Teste de Shapiro-Wilk (para cada grupo):

1. Ano 2015: Statistic=0.981, p-value=0.059
2. Ano 2014: Statistic=0.977, p-value=0.026
3. Ano 2013: Statistic=0.971, p-value=0.008
4. Ano 2012: Statistic=0.972, p-value=0.009
5. Ano 2011: Statistic=0.964, p-value=0.002
6. Ano 2010: Statistic=0.948, p-value=0.000
7. Ano 2009: Statistic=0.964, p-value=0.002
8. Ano 2008: Statistic=0.959, p-value=0.001
9. Ano 2007: Statistic=0.954, p-value=0.001
10. Ano 2006: Statistic=0.936, p-value=0.000
11. Ano 2005: Statistic=0.941, p-value=0.000
12. Ano 2004: Statistic=0.911, p-value=0.000
13. Ano 2003: Statistic=0.892, p-value=0.000
14. Ano 2002: Statistic=0.858, p-value=0.000
15. Ano 2001: Statistic=0.852, p-value=0.000

Esses resultados representam a análise de normalidade para cada ano. O teste de Shapiro-Wilk avalia se os dados de expectativa de vida em cada ano seguem uma distribuição normal. Com base nos resultados, podemos observar que, em muitos dos anos listados, os valores de p (p-value) são menores que 0,05, o que indica que os dados de expectativa de vida nesses anos

não seguem uma distribuição normal. Isso pode ser importante a considerar ao interpretar os resultados da ANOVA e ao escolher métodos estatísticos apropriados, já que a ANOVA pressupõe normalidade dos dados.

- Teste de Levene:

Statistic=0.829	p-value=0.637

O teste de Levene é usado para verificar a homogeneidade das variâncias entre os grupos, uma suposição importante para a análise de variância (ANOVA). Nesse caso, o valor P (0.637) é maior que 0,05. Portanto, com um nível de significância de 0,05, não há evidência estatística para rejeitar a hipótese nula. Isso sugere que a homogeneidade das variâncias entre os grupos é atendida, o que é uma boa notícia ao realizar uma análise de variância (ANOVA), pois a suposição de homogeneidade das variâncias é importante para a interpretação dos resultados da ANOVA.

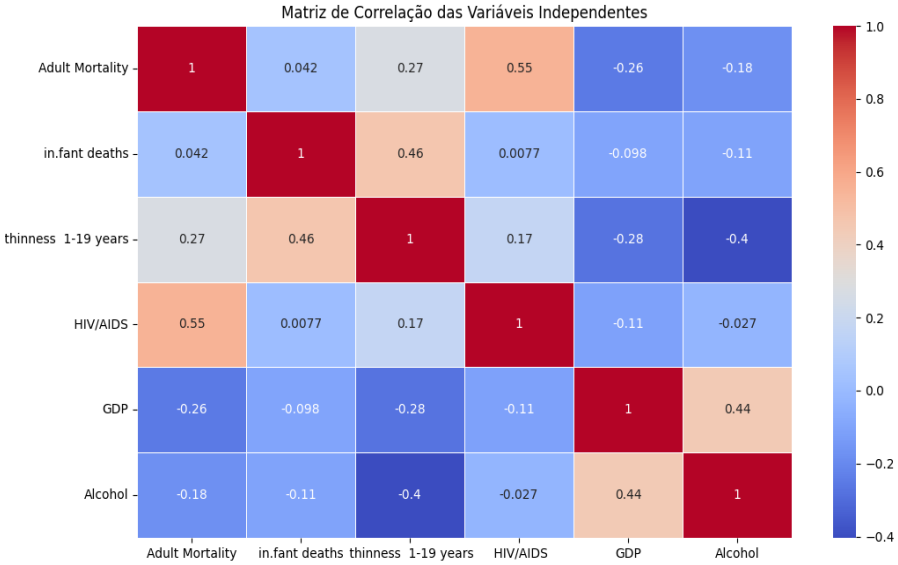
Em resumo, os resultados da ANOVA indicam que, globalmente, não há diferenças significativas na expectativa de vida entre os anos analisados. No entanto, ao examinar a normalidade dos dados, observamos que alguns anos apresentam distribuições não normais (com p-values menores que 0.05 nos testes de Shapiro-Wilk), o que pode afetar a robustez dos resultados da ANOVA. Apesar disso, o teste de Levene sugere que a homogeneidade das variâncias é atendida, o que é um pressuposto importante para a ANOVA.

Embora não haja diferenças significativas entre os anos em termos de expectativa de vida de acordo com a ANOVA global, é importante considerar a normalidade dos dados ao interpretar esses resultados. Alguns anos podem mostrar diferenças mais claras do que outros.

3.4 MULTICOLINEARIDADE E ANÁLISE DE RESÍDUOS

Multicolinearidade: A multicolinearidade é um fenômeno estatístico que ocorre quando duas ou mais variáveis independentes em um modelo de regressão estão altamente correlacionadas entre si. Isso pode criar problemas na interpretação dos resultados e na estabilidade das estimativas.

- Gráfico de correlação:



Com o gráfico podemos visualizar a relação entre as variáveis independentes, percebemos que não temos muitas relações fortes (mais de 0.5), somente uma, a relação de HIV com mortalidade de adultos, o restante das relações são intermediarias ou fracas, tendo até mesmo relações negativas, como HIV com Álcool.

- Análise VIF para multicolinearidade:

O VIF (Fator de Inflação da Variância) é uma métrica usada para avaliar a multicolinearidade em modelos de regressão linear múltipla. Ele indica o grau de multicolinearidade entre as variáveis independentes (também conhecidas como variáveis preditoras ou explanatórias) no modelo. Em essência, o VIF quantifica o quanto a variância da estimativa de um coeficiente de regressão é aumentada devido à multicolinearidade.

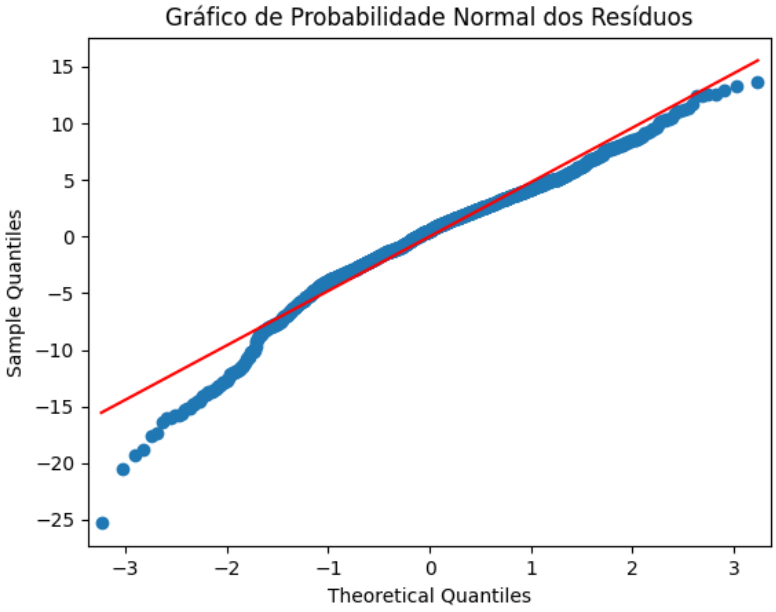
Multicolinearidade	**VIF**
Adult Mortality	3.148285
in.fant deaths	1.373477
thinness 1-19 years	2.515901
HIV/AIDS	1.498071
GDP	1.575365
Alcohol	2.101308

Geralmente, valores de VIF abaixo de 5 são considerados aceitáveis, indicando que a multicolinearidade não é um problema sério para as variáveis. Nesse caso, todos os VIFs estão abaixo desse limite, sugerindo que a multicolinearidade entre essas variáveis não é preocupante.

Análise de Resíduos: A análise de resíduos é uma etapa crítica na modelagem estatística, especialmente em modelos de regressão. Envolve a avaliação dos resíduos, que são as diferenças entre os valores observados e os valores previstos pelo modelo. Aqui estão alguns aspectos importantes:

- **Homocedasticidade**: A análise de resíduos ajuda a verificar se a variabilidade dos resíduos é constante em todas as faixas de valores previstos. A presença de heterocedasticidade (variabilidade não constante) pode indicar problemas na precisão do modelo.
- **Normalidade dos resíduos**: Os resíduos devem ser aproximadamente normalmente distribuídos. Desvios significativos da normalidade podem indicar que o modelo não se ajusta bem aos dados ou que transformações são necessárias.

Gráfico QQ-Plot dos resíduos - verifica a suposição de normalidade dos resíduos.



Como podemos ver, muitos pontos no gráfico QQ-Plot dos resíduos estão próximos a linha diagonal (ou seja, a linha de probabilidade normal), isso sugere que os resíduos do modelo estão próximos de uma distribuição normal, isso é uma boa notícia, pois é uma das suposições fundamentais para a validade de um modelo de regressão linear. apesar disso podemos ver que existem casos extremos, principalmente na parte inferior do gráfico.

- Testes para Verificar as Suposições:

Teste de Shapiro-Wilk:

Estatística de Teste:	0.9651119709014893
Valor-p:	1.6480186885123343e-19

Este teste avalia se os resíduos do modelo seguem uma distribuição normal. Um valor-p (p-value) baixo indica que os resíduos não seguem uma distribuição normal. No seu caso, o valor-p é baixo (1.6480186885123343e-19), o que sugere que os resíduos não seguem uma distribuição normal. Isso se deve ao fato mencionado acima, alguns valores dos resíduos têm valores extremos e discrepantes, prejudicando a normalidade dos resíduos.

Teste de White para Homocedasticidade:

Estatística de Teste:	427.5845587790964
Valor-p:	1.1721726383517959e-73

Este teste verifica se a variabilidade dos resíduos é constante em todos os níveis da variável independente. Um valor-p baixo indica que a homocedasticidade não é satisfeita. No seu caso, o valor-p é muito baixo (1.1721726383517959e-73), o que sugere que a homocedasticidade não é atendida. Provavelmente a falta de homocedasticidade se dá devido a presença de outliers (valores extremos) que podem aumentar a variabilidade dos erros e levar à falta de homoscedasticidade.

4. CONCLUSÃO

Após essa análise detalhada dos dados, obtivemos um entendimento mais profundo do conjunto de dados utilizado e fomos capazes de responder às perguntas propostas. Realizamos uma modelagem de regressão linear múltipla, que, apesar de várias transformações nas variáveis, resultou em um coeficiente de determinação (R^2) de 70%. Esse valor representa um resultado intermediário.

Ainda assim, não conseguimos alcançar um ajuste ótimo com um modelo de regressão linear. Isso sugere que o modelo linear pode não ser a escolha mais adequada para esses dados. Pode ser necessário realizar uma análise mais aprofundada do dataset e considerar a possibilidade de ajustar um modelo não linear, que seja capaz de capturar as relações mais complexas presentes nos dados. Portanto, futuras investigações podem se concentrar em explorar modelos mais complexos que melhor se adequem a essa complexidade subjacente dos dados

5. REFERÊNCIAS

FERREIRA, D.F. Estatística básica. UFLA, Lavras, 2009.

LAKSIKA THAMALINGAM. Kaggle. 2023. DISPONÍVEL EM: https://www.Kaggle.Com/datasets/uom190346a/health-and-demographics-dataset. Acesso em: 25/10/2023

MAGALHÃES, M. N.; LIMA, A. C. P. Noções de Probabilidade e Estatística. 4. ed. São Paulo: EDUSP, 2002.

MAGALHÃES, M. N.; LIMA, A. C. P. Noções de Probabilidade e Estatística. 4. ed. São Paulo: EDUSP, 2002.