Métodos Quantitativos

Lucas Esteves dos Reis Salgado, Rafaela Fernandes Horta
Universidade Federal de Juiz de Fora, Departamento de Ciências da Computação.
Rua José Lourenço Kelmer, s/n, Campus Universitário, Bairro São Pedro,
CEP:36036-900, Juiz de Fora-MG
lucas.esteves@engenharia.ufjf.br, rafaela.horta@engenharia.ufjf.br.

Resumo:
O objetivo deste trabalho é aplicar o conteúdo lecionado nas aulas de métodos quantitativos, examinando a estrutura e configuração dos dados, além de aprender sobre a relação entre as variáveis da base de dados. A Análise Exploratória de Dados inclui um conjunto de ferramentas descritivas e gráficas para buscar padrões e tendências que desempenharam o papel de hipóteses para uma análise completa. Empregando técnicas estatísticas descritivas e gráficas para estudar o conjunto de dados, detectando outliers e anomalias, e comunicando de forma eficaz os resultados do modelo.

1. INTRODUÇÃO

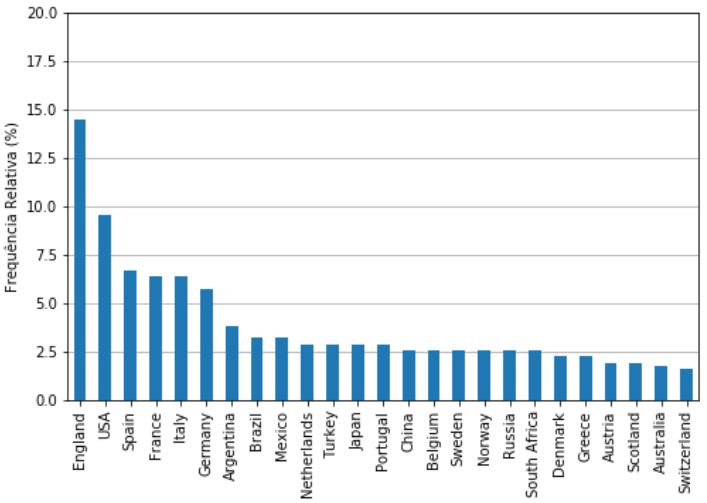
Utilizaremos o ambiente computacional do Jupyter Notebook, que nos permite importar diversas bibliotecas, como matplotlib, scipy, numpy, pandas, seaborn, entre outras, que serão usadas nesse estudo para coletar e tratar as informações.

2. ENCONTRANDO A FONTE DE DADOS

A primeira etapa para o data mining é encontrar nossa fonte de dados, para isso podemos buscar os dados já disponibilizados em plataformas como Kaggle, geralmente em formatos csv ou fazer o web scraping, que consiste em extrair de qualquer site o seu conteúdo para fazer uma análise.
Para fazermos web scraping é necessário buscar os dados pela ferramenta Inspecionar do navegador e utilizar a biblioteca do python urllib que faz a conexão com a página Web, também pela urllib é feita a verificação da conexão. Visto isso, é necessário a utilização da biblioteca BeautifulSoup para análise dos dados HTML e XML extraídos do website.
Para coletar a base de dados analisada realizamos o processo de web scrapping na seguinte url: https://projects.fivethirtyeight.com/global-club-soccer-rankings/ e baixamos a tabela que possui o ranking de 629 times internacionais de futebol que compara suas avaliações ofensivas, defensivas e força futebolística.

3. VARIÁVEIS DISCRETAS

As variáveis discretas analisadas foram os países e as ligas das quais os times fazem parte.
Plotando o gráfico da massa de probabilidades, em que a variável x representa todos os países da base de dados, geramos esse resultado:

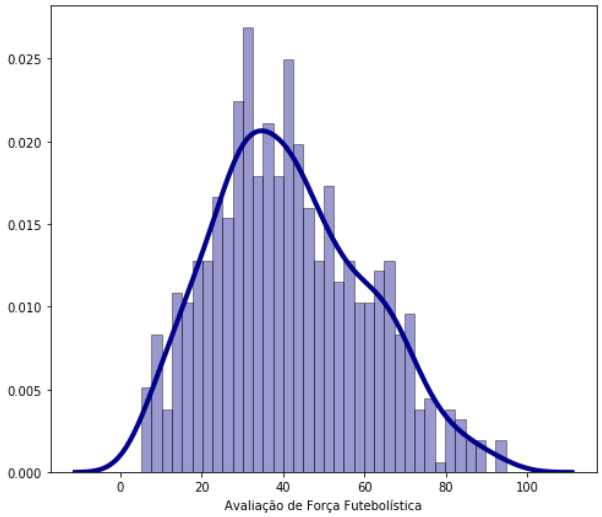


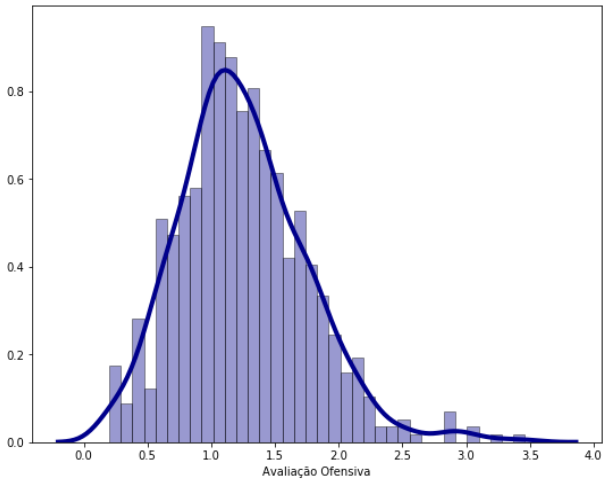
Podemos concluir que a maior parte dos times são da Inglaterra.

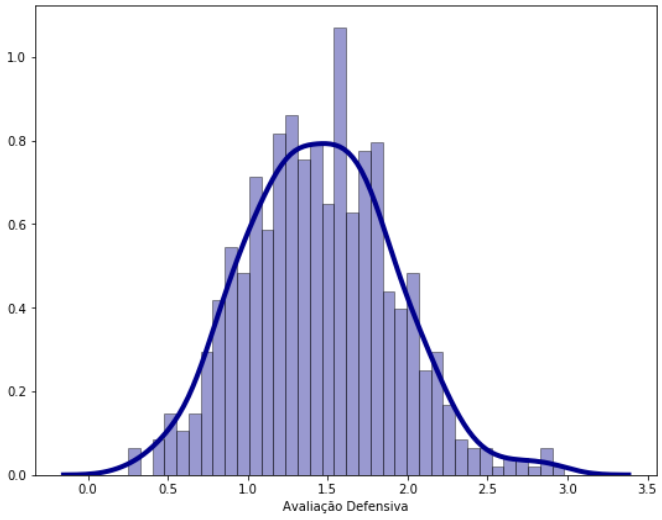
4. VARIÁVEIS CONTÍNUAS

Na amostra escolhida há algumas variáveis contínuas: a avaliação das ofensivas, que varia de 0.200000 à 3.460000, a avaliação da defensiva que varia de 0.250000 à 2.98000000 e a força futebolística, variando de 41.551574 à 94.740000.

Primeiramente utilizando o pacote seaborn plotamos os gráficos de densidade kernel para as três variáveis: avaliação de força futebolística, ofensiva e defensiva, respectivamente:

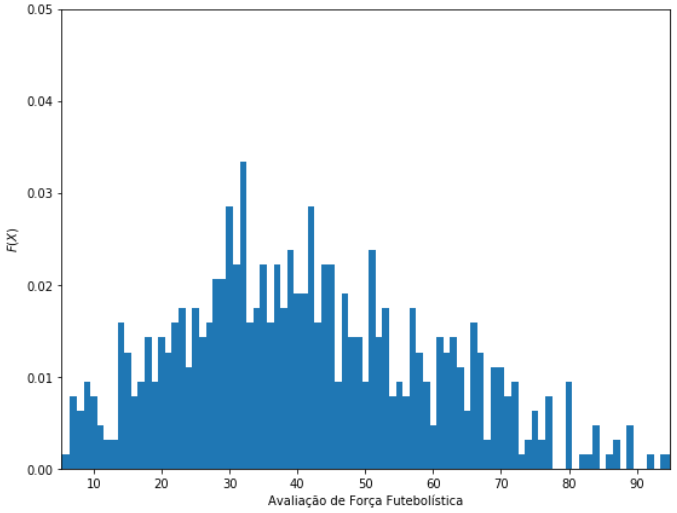


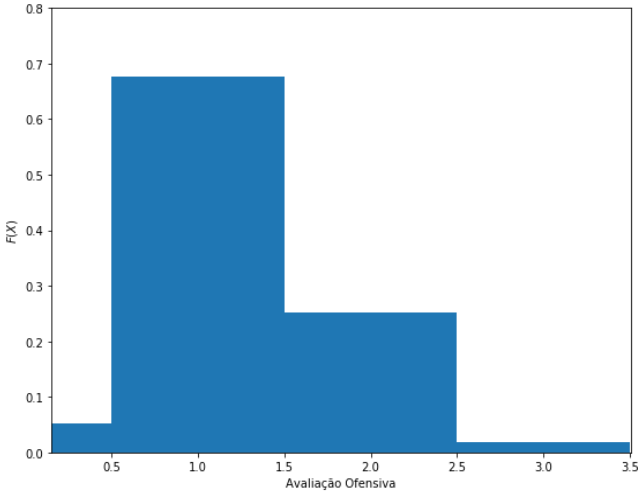


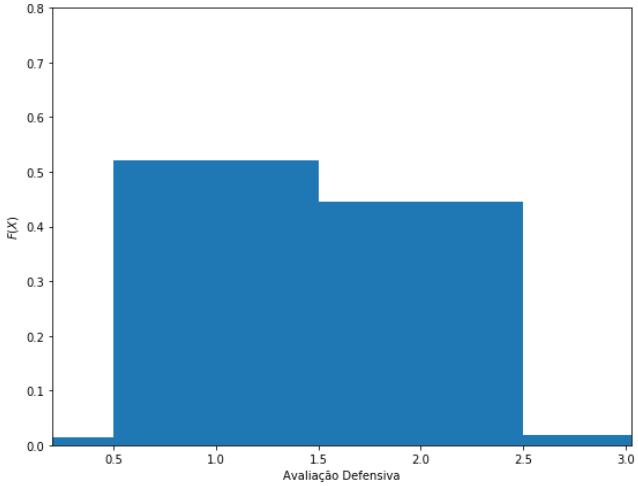


Nesse método, uma curva contínua (o kernel) é desenhada em todos os pontos de dados individuais e todas essas curvas são adicionadas juntas para fazer uma única estimativa de densidade suave. O kernel mais usado é um gaussiano (que produz uma curva de sino gaussiano em cada ponto de dados). O eixo x é o valor da variável como em um histograma. O eixo y é a função de densidade de probabilidade para a estimativa da densidade do kernel.

Também plotamos os gráficos de distribuição da densidade de probabilidade para as mesmas 3 variáveis:







Notamos pela distribuição da densidade que não é possível traçar uma curva normal para nenhuma das avaliações, mas mesmo assim iremos continuar as análises em busca de mais resultados.

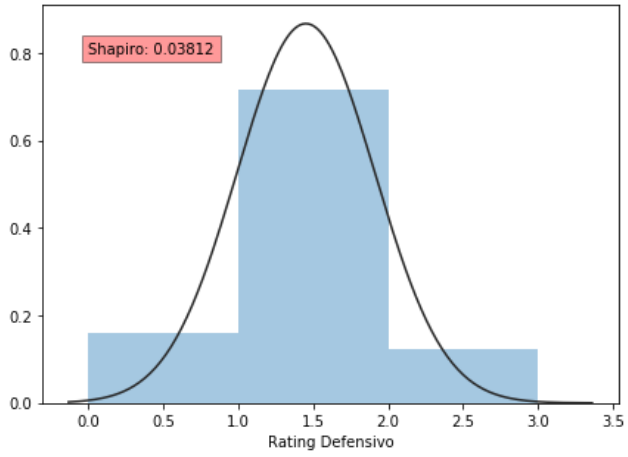
Para comprovar a hipótese nula, que afirma se a distribuição amostral é normal, aplica-se o teste Kolmogorov-Smirnov. Para tal, procuram-se p-values superiores a 0.8, conforme o especificado.
Usando a biblioteca scipy.stats encontramos os seguintes resultados:

- Avaliação de Força Futebolística:
  p-value= 0.0
- Avaliação Ofensiva:
  p-value=5.190505330200898e-265
- Avaliação Defensiva:
  p-value= 0.0

Os baixíssimos valores do p-value < 0.001 rejeitam a chance de constatação da Hipótese Nula, que afirma a normalidade da distribuição amostral.
É importante destacar que os valores p-value=0 não é uma verdade, mas uma aproximação, pois sempre há uma chance de obter os resultados da Hipótese Nula, por menor ou improvável que seja a chance.
É provável que a hipótese nula tenha sido rejeitada pelo tamanho da amostra.

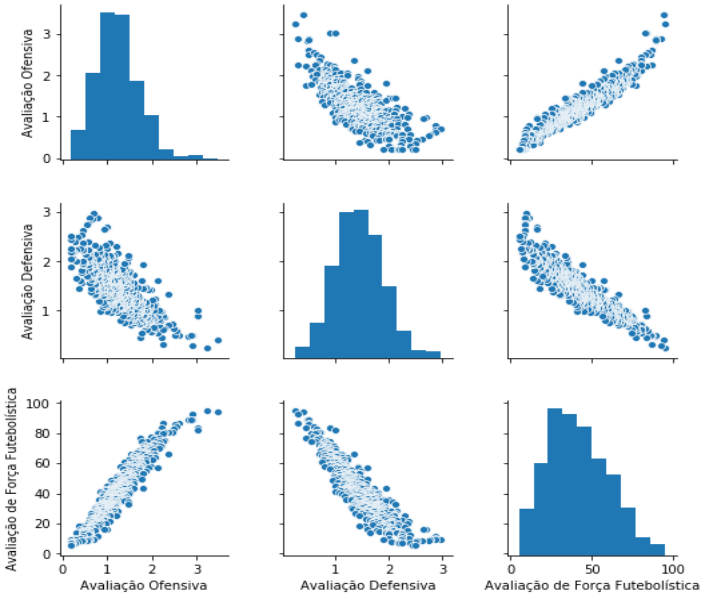
Tentaremos provar que há diferença entre os alvos da comparação estatística, confirmando a Hipótese Alternativa.

Realizamos o teste de Shapiro, que independe do tamanho da amostra, para encontrar um p-value significativo e a única amostra que apresentou mudança foi a da Avaliação Defensiva, porém ele continuou abaixo de 5% então rejeitamos novamente a hipótese nula.

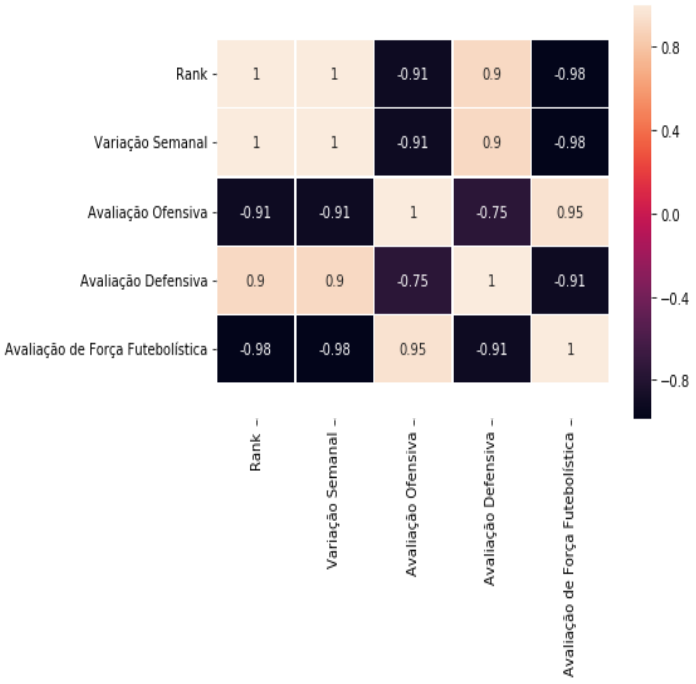


5. REGRESSÃO E PREDIÇÃO

Plotamos a matriz de dispersão abaixo e vemos que as diagonais mostram a distribuição de uma única variável em formato de histograma, enquanto as matrizes triangular inferior e superior mostram a relação entre duas variáveis.

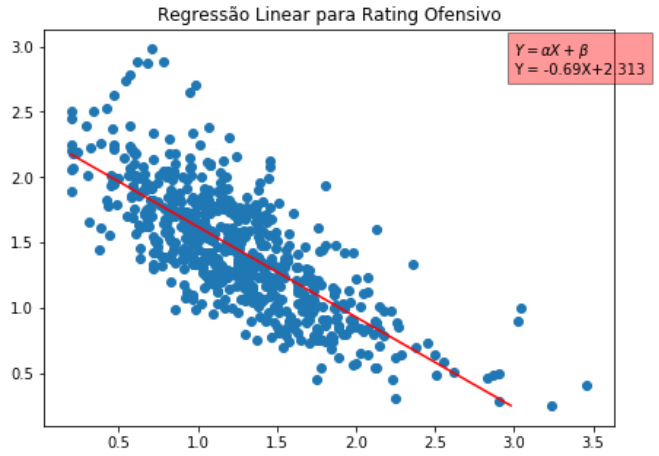


Em seguida plotamos o gráfico de correlação:

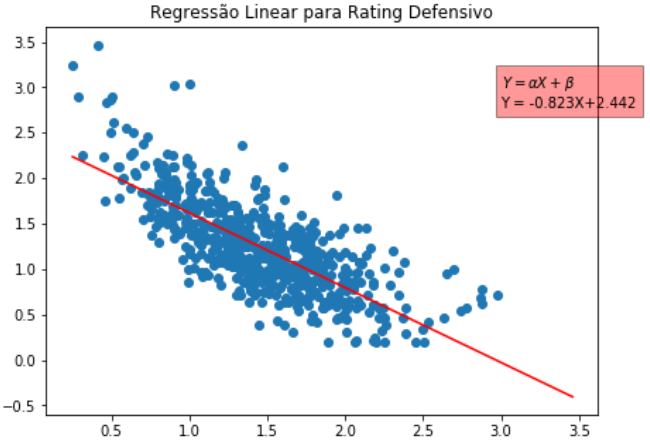


Como observado, a correlação para Avaliação Ofensiva e Defensiva com a Avaliação de Força Futebolística é muito forte porque os melhores times marcam muitos gols e sofrem poucos gols. Logo, nos melhores times, a Avaliação Ofensiva é alta, por isso há correlação positiva e a Avaliação Defensiva é baixa e por isso sua correlação negativa.
Podemos perceber que a correlação entre Avaliação Ofensiva (Gols marcados) x Avaliação Defensiva (Gols sofridos) é forte porque também é possível existirem times que fazem e sofrem muitos gols.
Calculamos a regressão de mínimos quadrados para dois conjuntos de medidas.

Primeiramente comparando a Avaliação Ofensiva x Avaliação Defensiva achamos os parâmetros: b0 = -0.6902839119853854 b1 = 2.3126571578637503, o valor para o teste de hipótese que a inclinação é nula: 2.210723813120407e-116, o coeficiente de determinação: 0.5680160212911692 e o desvio padrão da estimativa: 0.02404072212028615. O coeficiente de determinação não indica uma boa qualidade de regressão, porém o desvio padrão da estimativa indica uma baixa variabilidade. Pelo teste de hipótese não podemos considerar a inclinação nula.



Depois, comparando a Avaliação Defensiva x Avaliação Ofensiva achamos os parâmetros: b0 = -0.8228730402501329
b1 = 2.4416159063930865, o valor para o teste de hipótese que a inclinação é nula: 2.210723813120407e-116, o coeficiente de determinação: 0.5680160212911692 e o desvio padrão da estimativa: 0.028658442935502335. Como no conjunto anterior o coeficiente de determinação também não indica uma boa qualidade de regressão, mas o desvio padrão da estimativa indica uma baixa variabilidade. Pelo teste de hipótese não podemos considerar a inclinação nula.



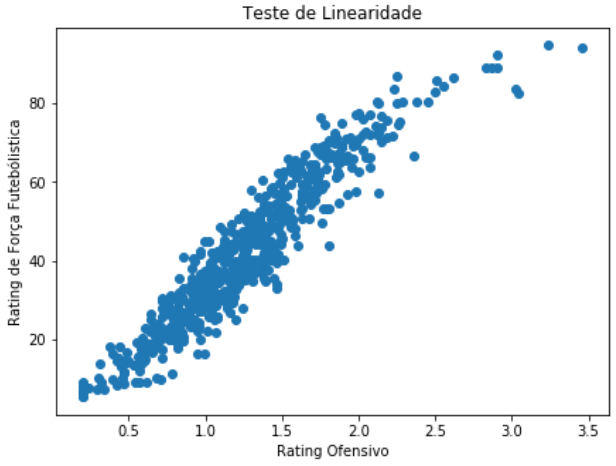
O conjunto Avaliação Ofensiva x Avaliação de Força Futebolística apresentou os parâmetros: b0 = 35.0707773859517, b1 = -2.1742698718360955, o valor para o teste de hipótese que a inclinação é nula: 4.792360331263e-310, o coeficiente de determinação: 0.8957601475585124 e o desvio padrão da estimativa: 0.47778515052136367. Foi o conjunto que indicou o melhor coeficiente de determinação e também uma boa qualidade de

regressão, com desvio padrão dos erros que indicam uma baixa variabilidade. (Pelo teste de hipótese não podemos considerar a inclinação nula.) Por esses melhores resultado usaremos esse conjuntos para fazer um estudo comparativo mais detalhado.

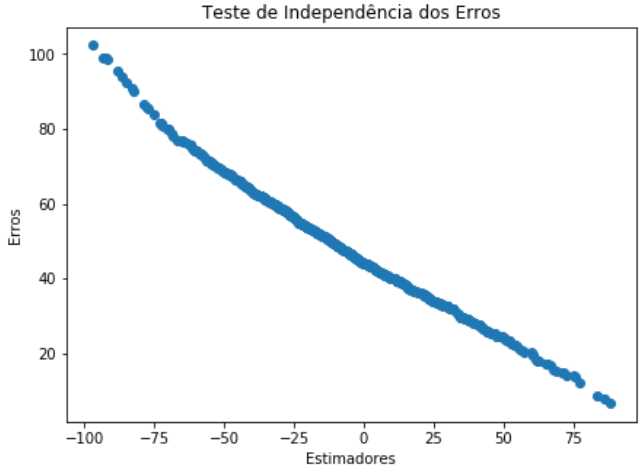
A equação de regressão é aproximadamente Y = 35.0708 + -2.1743x

Comprovando que R: 0.8957601475585101, concluímos que a qualidade da regressão é alta e possui desvio padrão dos erros: 6.021676069856358, desvio padrão do parâmetro b0:0.6422639398455074, desvio padrão do parâmetro b1:0.4777851505213685.

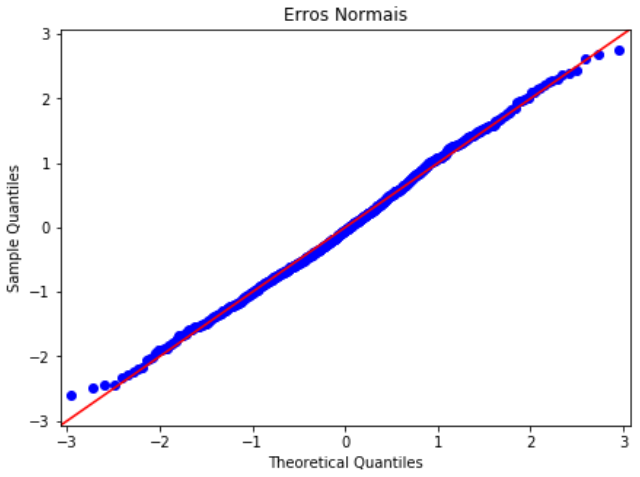
Para finalizar, realizamos os testes visuais de pressuposto



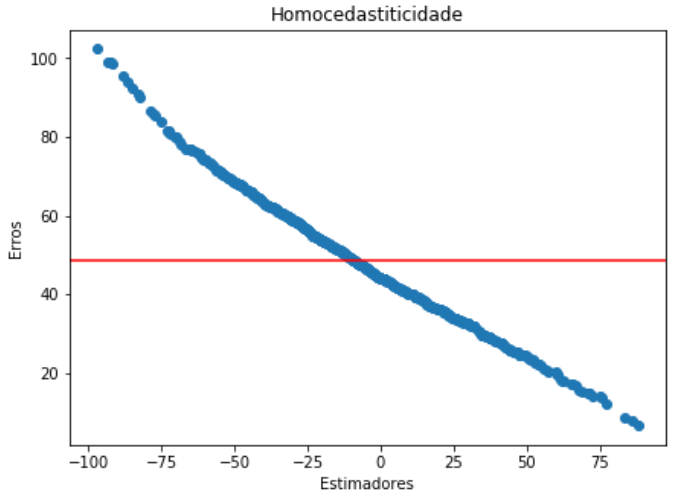
O teste de linearidade é bem válido se pensarmos em uma reta ajustada aos dados.



O erros aparentam seguir um padrão decrescente mostrando uma tendência visível, evidenciando uma dependência dos resíduos. Indica que um modelo de regressão não linear sobre a amostra pode apresentar melhores resultados.



Os erros estão seguindo uma distribuição normal, então podemos predizer e estimar em nosso modelo.



A distribuição dos dados em torno da média dos resíduos está com uma tendência visível, temos indícios que a variância dos resíduos não são homogêneas existindo heterocedasticidade. Essa tendência é um bom indício para uso de regressão não-linear.

6. CONCLUSÃO

Após todo o estudo da amostra concluímos que, como esperado, há forte correlação positiva entre a Avaliação Ofensiva x Avaliação de Força Futebolística e forte correlação negativa entre Avaliação Defensiva x Avaliação de Força Futebolística, ou seja os time que são mais fortes marcam mais gols e sofrem menos gols. Infelizmente a amostra não era normalizada e o modelo de regressão linear não se aplica adequadamente para ela.