**Relatório sobre Análise Exploratória de Dados (AED) de um Banco de Dados sobre a Flor Íris**

Checando primeiramente os dados de banco de dadosTela de celular

Descrição gerada automaticamente

Na primeira coluna (sepal\_lenght) temos os dados dos comprimentos das sépalas, posteriormente a largura das sépalas (sepal\_width) e em seguida o comprimento das pétalas (pental\_lenght) e depois a largura das pétalas (petal\_width) e por último a espécie de cada flor de íris (species). Cada linha representa um dado de cada flor.

Obtendo informações do DataFrame

Texto

Descrição gerada automaticamente

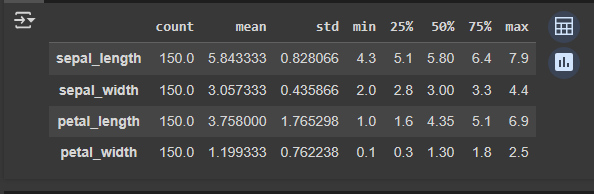
Possuem 150 entradas, que variam de 0 a 149. Temos 5 colunas, onde as 4 primeiras colunas são do tipo float, ou seja, números decimais e, por última coluna, é do tipo object que armazena um objeto do tipo de espécime e não temos nenhum valor vazio.

Verificando as espécime



O método iris.species.unique() nos permite saber quais espécies temos na última coluna

Verificando os dados estatísticos das colunas que possuem valores numéricos



O método iris.describe() retorna os valores das sépalas e pétalas, para afim de melhor visualização utilize também o comando .transpose().

Análise de Dados

Interface gráfica do usuário, Texto, chat ou mensagem de texto

Descrição gerada automaticamente

Podemos perceber que na média o comprimento das sépalas e pétalas é maior do que a largura, o que indica que os dois tipos de folhas modificadas são mais Compridos do que Largas. Portanto, temos uma evidência da característica da flor-de-íris.

Análise de Dados

Interface gráfica do usuário, Texto, Aplicativo

Descrição gerada automaticamente

Podemos observar que o desvio padrão das pétalas é maior do que das sépalas.

Análise de Dados

Interface gráfica do usuário, Texto, Aplicativo, chat ou mensagem de texto

Descrição gerada automaticamente

Podemos também observar que, quando comparando os comprimentos, vemos que as pétalas têm um desvio de padrão bem maior das sépalas e temos esse mesmo comportamento muito similar quando a gente observa a largura, isso indica que as pétalas em geral variam mais em tamanho do que as sépalas.

Análise de Dados

Interface gráfica do usuário, Texto

Descrição gerada automaticamente

Através dos valores mínimos, percentis e valores máximos podemos confirmar essa informação.

Análise de Dados com o Boxplot

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Temos uma informação gritante de outliers que foram identificados na largura das sépalas, que estão sendo representados por esses pontos em círculos destacados por essas setas em vermelho. Podendo não ser necessariamente outliers porque o boxplot representa um conjunto de dados dentro do intervalo de noventa e nove (99%) de confiança, então quando temos esses pontos marcados, como nesse caso, indica que estão fora do esperado para essa amostra. Caso tivéssemos acesso aos dados verdadeiros, teríamos que investigar o motivo desses outliers para decidir se mantemos ou não esses dados nos conjuntos de dados, ou não. Nesse caso, não irei remover nenhum dado.

Análise de Dados com o Boxplot

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

Outro dado que chama atenção é o tamanho do intervalo dos dados do comprimento das pétalas em relação aos demais. Já vimos isso anteriormente, porém, com a visualização do Boxplot podemos perceber isso mais evidentemente.

Análise de Dados com o Histograma

Texto

Descrição gerada automaticamente

A ideia da visualização em gráfico de Histograma é ter um pouco mais de clareza sobre como os dados estão sendo distribuídos.

Análise de Dados com o Histograma

Gráfico, Histograma

Descrição gerada automaticamente

Chama atenção que os dados não aparecem ser distribuídos de forma normal, na realidade a única distribuição que aparenta ser normal é a largura das sépalas.

Análise de Dados com o Histograma

Gráfico, Histograma

Descrição gerada automaticamente

Portanto, podemos perceber que a única que possui uma distribuição normal é a largura das sépalas, que lembra um formato de sino.

Análise de Dados com o Histograma

Gráfico, Histograma

Descrição gerada automaticamente

E talvez o comprimento das sépalas também seja uma distribuição normal, no entanto, não consigo ter essa confirmação apenas com análise visual.

Análise de Dados com o Histograma

Gráfico, Histograma

Descrição gerada automaticamente

Quando olhamos para a distribuição de dados das pétalas, parece ter pelo menos dois conjuntos de dados distintos nessa mesma distribuição, um que apresenta uma medida bem pequena e outro com medidas bem maiores.

Análise de Dados com o Histograma

Gráfico

Descrição gerada automaticamente

Isso fica mais claro quando olhamos para o boxplot do comprimento das pétalas, que é a mesma característica que vimos que tem maior dispersão em torno da sua média, ou seja, muito provavelmente temos um espécime da flor de íris que tem as pétalas bem menores do que as outras.

Checando os dados da distribuição

Texto

Descrição gerada automaticamente

Esse método verifica o gráfico histograma de outra forma, primeira eu passo o dado separadamente nesse caso foi comprimento das sépalas e modelo vai ser ajustado aos dados reais e nesse caso é modelo da distribuição normal e o modelo é encontrado com comando da biblioteca ‘stats.norm’ e posteriormente passamos o parâmetro ‘kde’ como ‘False’, pois, não queremos a densidade do calciano dos dados e depois basta pedir para mostrar o gráfico.

Checando os dados da distribuição

Gráfico, Histograma

Descrição gerada automaticamente

Aqui temos um histograma para nossa distribuição normal, esperando para nosso conjunto de dados dos comprimentos das sépalas. Mesmo com a distribuição normal ajustada, ainda é complicado dizer se os dados são normais ou se eles não de forma apenas visual.

Teste de Normalidade de Shapiro-Wilk

Texto

Descrição gerada automaticamente com confiança média

Como não consigo confirmar de forma visual o histograma anterior utilizei o teste de normalidade de Shapiro-Wilk.

O primeiro comando é passar a biblioteca stats, depois o método Shapiro junto com conjunto de dados ao qual queremos avaliar (nesse caso é o ‘sepal\_lenght’), portanto, estamos passando os dados da coluna ‘sepal\_lenght’ para o Shapiro.

Com o resultado o teste de Shapiro-Wilk retornar uma tupla, onde o primeiro elemento é estatístico (representado por vermelho em 1 em romano) do teste e segundo o valor de p (representado por roxo em 2 em romano)

Teste de Normalidade de Shapiro-Wilk

Texto

Descrição gerada automaticamente

Aqui criei duas variáveis, 1º para armazenar o valor estatístico (shapiro\_stat) e 2º para armazenar o valor de p (shapiro\_p). Posteriormente, para verificar se fiz correto, pedi para Colab imprimir o valor da variável de p.

Teste de Normalidade de Shapiro-Wilk

Texto

Descrição gerada automaticamente

Estou comparando o valor de p com o nível de tolerância pré-estabelecido (o qual passei o valor 0,05 que é um número altamente recomendado), portanto, criei um condicional (if) que se o valor de p for maior que o nível de tolerância irá emitir um aviso que a distribuição é normal, caso contrário irá emitir um aviso que a distribuição não é normal

Com resultado “A distribuição não é normal” podemos afirmar que comprimento das flores de íris não segue a distribuição normal

Gráfico de Dispersão

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

A melhor forma para verificar se existe algum tipo de padrão entre dois conjuntos de dados é por meio dos gráficos de dispersão, que trazem a relação entre duas variáveis de forma simples e objetiva.

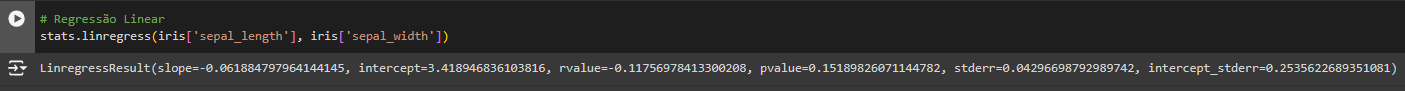
Gráfico de Dispersão

Gráfico, Gráfico de dispersão

Descrição gerada automaticamente

Podemos verificar se há uma correlação entre os valores que estão no eixo X e os valores do eixo Y. Em outras palavras, vamos avaliar o que acontece com a largura das sépalas quando temos flores que apresentam sépalas mais compridas.

Regressão Linear



A forma mais fácil de se fazer isso é com ajustar o modelo linear ao conjunto de dados que vai nos dizer qual é o nível de relação entre X e Y, para isso usamos a função linregress da biblioteca scipy.stats

Legenda:

* Slope é coeficiente da reta
* Intercept é coeficiente linear da reta
* Rvalue é o r de Pearson
* Pvalue é o valor de p desse ajuste
* Stderr é variância do modelo
* intercept\_stderr é o erro padrão do intercepto estimado na regressão linear

Com os atributos ‘slope’ e ‘intercept’ é possível criar uma reta no gráfico