Análise Exploratória de Dados - Curso

1. **O que é análise de dados?**

   A Análise Exploratória de Dados, antigamente chamada apenas de Estatística Descritiva, constitui o que a maioria das pessoas entende como Estatística, e inconscientemente usa no dia a dia.

   Consiste em RESUMIR E ORGANIZAR os dados coletados através de tabelas, gráficos ou medidas numéricas, e a partir dos dados resumidos procurar alguma regularidade ou padrão nas observações (INTERPRETAR os dados).

   A partir dessa interpretação inicial é possível identificar se os dados seguem algum modelo conhecido, que permita estudar o fenômeno sob análise, ou se é necessário sugerir um novo modelo.

2. **Objetivo**

   O objetivo da análise de dados é extrair informações significativas e úteis a partir de conjuntos de dados. Isso envolve examinar, limpar, transformar e modelar dados para descobrir padrões, tendências, relações e insights que possam orientar a tomada de decisões informadas. A análise de dados é aplicada em diversos campos, como negócios, ciência, saúde, finanças, governo e muitos outros.

   Alguns dos objetivos específicos da análise de dados incluem:

   Tomada de Decisões Informadas: A análise de dados ajuda a embasar decisões através de evidências quantitativas, reduzindo a incerteza e o risco associados a decisões baseadas apenas em intuição.

   Identificação de Padrões e Tendências: Permite a identificação de padrões repetitivos, tendências emergentes e anomalias nos dados, fornecendo insights valiosos.

Otimização de Processos: Ajuda a melhorar eficiências e identificar áreas de melhoria nos processos organizacionais, aumentando a eficácia e reduzindo custos.

Previsão e Antecipação: Por meio de técnicas como modelagem estatística e aprendizado de máquina, a análise de dados pode ser usada para prever resultados futuros com base em padrões históricos.

3. Quais os tipos de análise de dados?

3.1 Descritivo

Uma análise descritiva dos dados procura resumir as medições em um único conjunto de dados sem interpretação adicional. Um exemplo é o Censo. O Censo coleta dados sobre tipo de residência, localização, idade, sexo e raça de todas as pessoas nos Estados Unidos em um horário fixo. O Censo é descritivo porque o objetivo é resumir as medições neste conjunto fixo de dados em contagens populacionais e descrever. A interpretação e o uso dessas contagens são deixados ao Congresso e ao público, mas não fazem parte da análise dos dados. Calcular a média, mediana e moda de uma amostra de dados de altura de estudantes de uma escola para entender as tendências centrais e a distribuição dessa variável.

3.2 Exploratório

Uma análise exploratória de dados baseia-se em uma análise descritiva, buscando descobertas, tendências, correlações ou relações entre as medições de múltiplos atributos para gerar ideias ou hipóteses. Um exemplo é Utilizar gráficos de dispersão, box plots e histogramas para explorar a relação entre as horas de estudo e as notas de estudantes, identificando padrões e possíveis outliers.

3.3 Inferencial

Uma análise inferencial de dados vai além de uma análise exploratória, quantificando se um padrão observado provavelmente se manterá além do conjunto de dados em questão. As análises inferenciais de dados são as análises estatísticas mais comuns na literatura científica formal. O objetivo é identificar a força do relacionamento tanto no conjunto de dados específico quanto determinar se esse relacionamento se manterá em dados futuros.

3.4 Preditivo

Uma análise preditiva de dados utiliza um subconjunto de medições (as características) para prever outra medição (o resultado). Um exemplo é quando organizações usam dados de pesquisas para prever como as pessoas votarão no dia das eleições. Em alguns casos, o conjunto de medidas utilizadas para prever o resultado será intuitivo. Mas as análises preditivas de dados mostram apenas que é possível prever uma medição a partir de outra, mas não explicam necessariamente por que essa escolha de previsão funciona.

4. Bibliotecas para análise de dados Python

Em resumo, as bibliotecas Python reúnem diversas funções cujo objetivo é reduzir o uso de código no programa. Entretanto, quando falamos de bibliotecas Python para análise de dados o papel delas é fazer o tratamento e análise dos dados, criando assim modelos preditivos e possibilitando a mineração dos dados.

Abaixo, falamos das bibliotecas mais indicadas para análise de dados Python:

1. Pandas

Uma das principais bibliotecas da linguagem Python, a Pandas é muito utilizada para executar operações de dados em alta performance. Por exemplo, para análise ou manipulação dos dados.

De modo geral, a Pandas trabalha com duas estruturas: Dataframes, estruturas bidimensionais de dados que se assemelham a uma planilha

do Excel, e Series, que podem ser definidos como um array unidimensional ou uma lista simples.

Pandas é uma biblioteca que aceita arquivos em diferentes formatos, como csv e xlsx. Além disso, com ela é possível realizar operações de álgebra relacional, fazer o preenchimento ou a substituição de valores nulos, entre outras operações.

Por todos esses motivos, a Pandas é uma das bibliotecas mais utilizadas por Cientistas de Dados.

2. IPython

O IPython é uma biblioteca integrada ao Projeto Jupyter onde a biblioteca pode ser utilizada para fins diversos. Por exemplo, para desenvolvimento de pequenos programas em Python, para visualização de dados ou como ferramentas de cálculo. Nessa biblioteca diversas operações numéricas estão disponíveis, como adição, subtração, multiplicação, divisão, entre outros.

As bibliotecas para análise de Dados Python permite o tratamento e análise em um só lugar.

3. NumPy

Em resumo, podemos definir o Numpy como uma biblioteca Python que é muito utilizada para o cálculo de arrays e matrizes multidimensionais. Além disso, ela é uma ferramenta bem completa com suporte às diversas funções de álgebra linear e capacidade de integração a outras ferramentas.

4. Matplotlib

Matplotlib é uma biblioteca voltada para a visualização de dados e a criação de gráficos, como o de duas dimensões, com eixo x e y, ou os gráficos de calor. Matplotlib é uma biblioteca de código aberto, totalmente gratuita.

5. Seaborn

A Seaborn é uma ferramenta construída com base na biblioteca Matplotlib. Isso significa que, assim como a Matplotlib, Seaborn é uma ferramenta orientada a gráficos estatísticos. Entretanto, o seu diferencial é a aparência dos gráficos, que se tornam visualmente mais agradáveis devido às opções de personalização, como paleta de cores e fundo do gráfico.

6. SciPy

O SciPy (https://scipy.org/) é uma coleção de pacotes voltada para uma série de diversos domínios de problemas padrões no processamento científico.

7. Scikit-learn

Desde a criação do projeto em 2010, o scikit-learn (http://scikitlearn.org/stable/) se transformou no principal kit de ferramentas de propósito geral para aprendizado de máquina dos programadores Python. Em apenas sete anos, teve mais de 1.500 colaboradores em todo o mundo. O scikit-learn inclui submódulos para módulos como:

• Classificação: SVM, vizinhos mais próximos (nearest neighbors), floresta aleatória (random forest), regressão logística etc.

• Regressão: regressão de Lasso, regressão de ridge etc.

• Clustering: k-means, clustering espectral etc.

• Redução de dimensionalidade: PCA, seleção de atributos, fatoração de matrizes etc.

• Seleção de modelos: grid search (busca em grade), validação cruzada, métricas.

• Pré-processamento: extração de atributos, normalização. Junto com o pandas, o statsmodels e o IPython, o scikit-learn tem sido crucial para possibilitar que Python seja uma linguagem de programação produtiva para a ciência de dados.

5. **O que é um dataset?**

Conjunto estruturado de informações que são organizadas e armazenadas para serem analisadas. Cada conjunto de dados consiste em uma coleção de entradas individuais, onde cada entrada geralmente representa uma observação ou uma instância. Essas entradas são organizadas em colunas (também conhecidas como variáveis ou atributos) que representam diferentes características ou propriedades das observações.

Exemplo:

O conjunto de dados Iris é considerado o Hello World para ciência de dados. Ele contém cinco colunas: comprimento da sépala, largura da sépala, comprimento da pétala, largura da pétala e tipo de espécie. A íris é uma planta com flores, os pesquisadores mediram várias características das diferentes flores da íris e as registraram digitalmente.

	sepallength	sepalwidth	petallength	petalwidth	class
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa
...	...	...	...	...	...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

6. **Atributos**

Quando um determinado fenômeno é estudado determinadas características são analisadas: os atributos. É através dos atributos que se torna possível descrever o fenômeno. Os atributos são características que podem ser observadas ou medidas em cada elemento

pesquisado (seja por censo ou amostragem, levantamento ou experimento), sob as mesmas condições. Para cada atributo, para cada elemento pesquisado, em um dado momento, há um e apenas um resultado possível. Os atributos podem basicamente ser classificados de acordo com o seu nível de mensuração (o quanto de informação cada atributo apresenta) e seu nível de manipulação.

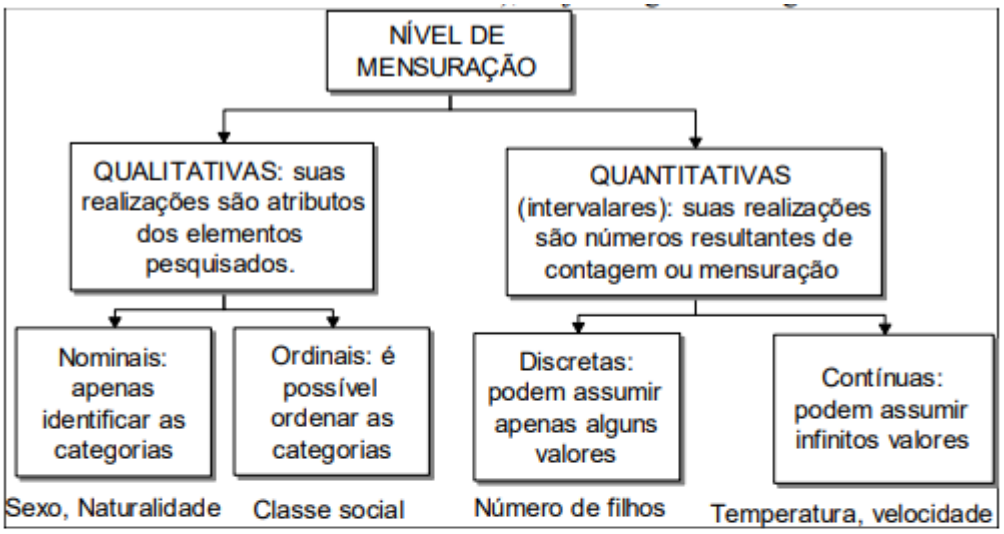


Figura 5 - Classificação das variáveis por nível de mensuração

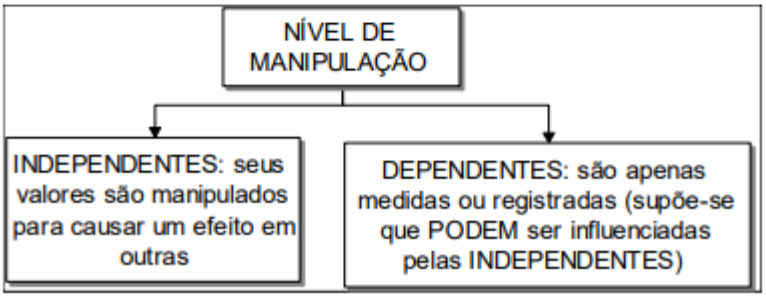


Figura 6 - Classificação das variáveis por nível de manipulação

6.1 - Classificação por nível de mensuração

A primeira classificação divide os atributos em QUALITATIVOS e QUANTITATIVAS. **Os atributos QUALITATIVOS ou categóricos são aqueles cujas realizações são atributos (categorias) do elemento pesquisado, como o sexo, grau de instrução, espécie. Os atributos QUALITATIVOS podem ser NOMINAIS ou ORDINAIS.** Os atributos NOMINAIS podem ser medidos apenas em termos de quais itens pertencem a diferentes categorias, mas não se pode quantificar nem mesmo ordenar tais categorias. Por exemplo, pode-se dizer que 2 indivíduos são diferentes em termos da atributo A (sexo, por exemplo),

mas não se pode dizer qual deles “tem mais” da qualidade representada pela atributo.

Exemplos típicos de variáveis nominais são sexo, naturalidade, etc. os atributos ORDINAIS permitem ordenar os itens medidos em termos de qual tem menos e qual tem mais da qualidade representada pelo atributo, mas ainda não permitem que se diga “o quanto mais”. Um exemplo típico de uma atributo ordinal é o status socioeconômico das famílias residentes em uma localidade: sabe-se que média alta é mais “alta” do que média, mas não se pode dizer, por exemplo, que é 18% mais alta.

Já **os atributos QUANTITATIVOS ou numéricos são aqueles cujas realizações são números resultantes de contagem ou mensuração, como número de filhos, número de clientes, velocidade em km/h, peso em kg, etc. Os atributos quantitativos também costumam ser divididos em DISCRETOS e CONTÍNUOS.** Os atributos QUANTITATIVOS DISCRETOS são aquelas que podem assumir apenas alguns valores numéricos que geralmente podem ser listados (número de filhos, número de acidentes). Os atributos QUANTITATIVOS CONTÍNUOS são aqueles que podem assumir teoricamente qualquer valor em um intervalo (velocidade, peso).

A predileção dos pesquisadores em geral por variáveis quantitativas explica-se porque elas costumam conter mais informação do que as QUALITATIVOS. Quando o atributo peso de um indivíduo é descrito em termos de “magro” e “gordo” sabemos que o gordo é mais pesado do que o magro, mas não temos idéia de quão mais pesado. Se, contudo, descreve-se o peso de forma numérica, medido em quilogramas, e um indivíduo pesa 60 kg e outro pesa 90 kg, não somente sabemos que o segundo é mais pesado, mas que é 30 kg mais pesado do que o primeiro.

É importante ressaltar que a forma como a atributo está sendo medido definirá o seu nível de mensuração. Por exemplo, a atributo velocidade de um carro. Se definirmos velocidade como resultado de uma medição por meio de radar resultando em um valor em km/h trata-se de uma atributo quantitativo contínuo. Se, porém, definirmos a velocidade como resultado de uma medição em que alguém declara a velocidade como "baixa", "média" ou "alta", ele passa a ser qualitativo ordinal.

6.2 - Classificação pelo nível de manipulação

Outra forma de classificar os atributos refere-se à sua manipulação: atributos INDEPENDENTES e DEPENDENTES.
Atributos INDEPENDENTES são aqueles que são manipulados enquanto que atributos DEPENDENTES são apenas medidos ou registrados (como resultado da manipulação dos atributos independentes). Esta distinção confunde muitas pessoas que dizem que “todos os atributos dependem de alguma coisa”. Entretanto, uma vez que se esteja acostumado a esta distinção ela se torna indispensável.

"Os atributos independentes são aquelas que PODEM INFLUENCIAR os valores dos atributos dependentes".

Os termos atributo dependente e independente aplicam-se principalmente à pesquisa experimental, onde alguns atributos são manipulados, e, neste sentido, são “independentes” dos padrões de reação inicial, intenções e características das unidades experimentais. Espera-se que outros atributos sejam “dependentes” da manipulação ou das condições experimentais. Ou seja, elas dependem “do que as unidades experimentais farão” em resposta.

Contrariando um pouco a natureza da distinção, esses termos também são usados em estudos em que não se manipulam atributos independentes, literalmente falando, mas apenas se designam sujeitos a “grupos experimentais” (blocos) baseados em propriedades pré-existentes dos próprios sujeitos.

- ***Os dados podem ser usados para responder a muitas perguntas, mas não a todas****. Um dos cientistas de dados mais inovadores de todos os tempos disse isso melhor. Os dados podem não conter a resposta. A combinação de alguns dados e um desejo ardente por uma resposta não garante que uma resposta razoável possa ser extraída de um determinado conjunto de dados. Antes de realizar uma análise de dados, o segredo é definir o tipo de pergunta que será feita. Algumas perguntas são mais fáceis de responder com dados e outras são mais difíceis. Esta é uma categorização ampla dos tipos de*

questões de análise de dados, classificadas de acordo com a facilidade de resposta à questão com dados

7. *Etapas de Análise*

De acordo com AZEVEDO e SANTOS (2008), o processo de descoberta de conhecimento em bases de dados é um processo interativo e iterativo, que apresenta as seguintes etapas: a) Seleção; b) Pré-processamento; c) Transformação; d) Data Mining; e) Interpretação, três desses se referem a nossa etapa de exploração e análise de dados, falaremos deles.

7.1 Seleção

Primeiramente, é necessário desenvolver a compreensão do domínio de aplicação, o conhecimento relevante que pode ser extraído e os objetivos do processo segundo a "visão do cliente", reduzindo, assim, o contexto da exploração das informações. Logo em seguida, deve-se definir o conjunto de dados alvo ou subconjunto de variáveis ou amostra de dados que proporcionará a descoberta de conhecimento útil. ***Nessa etapa, um fator importante é determinar um conjunto de dados relevante que possivelmente descreverá melhor os dados para que conhecimento seja adquirido posteriormente (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996)****.*

7.2 - Pré-análise dos Dados

Corresponde à etapa de pré-tratamento dos dados, removendo ruídos, quando possível, ou adquirindo informações para conseguir lidar com eles. *Também engloba decisões estratégicas para lidar com dados incompletos, controlar informações que variam com o tempo e possíveis mudanças que podem ocorrer (FAYYAD, PIATETSKY-SHAPIRO e SMYTH, 1996).*

7.3 - Transformação

A transformação dos dados pode ser determinada pela redução à dados relevantes para a pesquisa, assim como a projeção desses

dados numa visão futura, determinando características relevantes dos dados de acordo com o objetivo. *Pode-se utilizar métodos de redução ou transformação dimensional para determinar um subconjunto de variáveis consideráveis para o problema. Tendo em vista a praticidade do processo, é possível reduzir conjuntos de variáveis para facilitar as etapas seguintes sem que possivelmente, se perca informação relevante.*

8. ***Engenharia de Features***

A engenharia de features é o que vai ensinar como ter um aprendizado de máquina eficiente. *Assim, de forma bem direta. E é dessa forma direta que faço a seguinte pergunta: quanto mais features, melhor?*

Antes de dar um overview sobre a engenharia de features, é importante entendermos algumas definições. Suponhamos que se tenha dados brutos tabulados – ou em um banco de dados – para servir de input para uma análise de um treinamento de aprendizado.

Estes dados tabulados seriam compostos por linhas e colunas, onde as colunas representariam os atributos, e as linhas as instâncias – ou seja, a quantidade de exemplos para o conjunto de atributos.

Os atributos portanto, representam a propriedade de um objeto e definem um valor específico para uma certa instância. Só que nem todos os atributos devem estar presentes nesse conjunto de dados, ou dataset.

Por exemplo: para se fazer uma previsão de ocorrência de acidente de trânsito, baseada nos atributos dos clientes, podemos perceber que os atributos CPF e RG não fazem diferença, e na verdade podem até atrapalhar a construção do aprendizado de uma rede neural. Neste sentido, utilizamos a palavra feature para definir um atributo que tenha mais importância e mais significado.

Então, respondendo a pergunta acima: não necessariamente. Mais features precisam de maior monitoramento e com isso mais engenharia.

É e dessa forma que começamos a falar sobre o processo de engenharia de features, que tem por objetivo filtrar ainda mais essas features. Existem casos em que necessitamos expandir, selecionar ou excluir algumas delas. Como a feature idade, que pode ser excluída se tivermos a feature data de nascimento, já que elas são redundantes e o modelo pode não convergir tão bem.

Agora, vamos por partes. Separei aqui os principais tópicos para discutirmos sobre engenharia de features e dados.

O que são features?

Como já falamos acima, Features ou, em português, características, são fatores utilizados para definir um atributo que tenha mais importância e significado.

O trabalho de engenharia de features consiste em elencar as características mais importantes dos dados e, dependendo do modelo utilizado, normalizar os valores contidos na base de dados.

O processo de engenharia de features é a fase mais importante e trabalhosa quando se deseja ter um aprendizado de máquina eficiente. Se, por exemplo, a modelagem estiver razoável, porém, com excelente tratamento dos dados, teremos um melhor resultado do que com uma modelagem excelente e dados com uma preparação mais pobre, seja por inconsistência, ruídos, dados faltantes ou até mesmo pouca quantidade de informação.

Podemos dizer então que é um processo iterativo para transformar dados brutos em características que melhor representam o problema. Um caso muito comum disso são os modelos de deep learning que recebem dados brutos e internamente já identificam as melhores features como as bordas, linhas e texturas em uma classificação de imagem.

Seleção de Variáveis:

A seleção de variáveis é uma parte crucial da engenharia de features, pois visa identificar e manter apenas as características mais relevantes para o problema em questão. Existem várias técnicas para isso, como:

-Importância das Variáveis:

Utilizar métricas como a importância das variáveis provenientes de modelos, como árvores de decisão, para identificar quais características mais contribuem para a predição.

-Análise Estatística:

Aplicar técnicas estatísticas para entender a correlação entre variáveis e eliminar aquelas que têm baixa influência ou que são altamente correlacionadas.

-Regularização:

Algoritmos de regularização, como L1 (Lasso) e L2 (Ridge), podem ser empregados para penalizar coeficientes menos importantes, tornando alguns deles extrativos e, portanto, eliminando variáveis menos relevantes.

Criação de Novos Atributos:

A criação de novos atributos é outra estratégia importante na engenharia de features, pois pode melhorar a capacidade do modelo de capturar padrões. Algumas abordagens incluem:

- ***Engenharia de Polinômios:***

Elevar variáveis a potências mais altas para capturar relações não lineares.

- ***Combinação de Atributos:***

Criar novas variáveis combinando características existentes, o que pode revelar informações importantes para o modelo.

- ***Extração de Informação Temporal:***

Para conjuntos de dados temporais, extrair características como sazonalidade, tendências e ciclos pode ser valioso.

Transformação de Variáveis:

A transformação de variáveis é essencial para garantir que os dados estejam em uma forma que o modelo possa entender e aprender eficientemente. Algumas técnicas comuns incluem:

- ***Normalização e Padronização:***

Garantir que as variáveis estejam na mesma escala, evitando que uma característica com grande amplitude domine outras.

- ***Transformações Logarítmicas ou Exponenciais:***

Aplicar transformações para lidar com distribuições assimétricas e reduzir o impacto de outliers.

Tratamento de Variáveis:

O tratamento de variáveis envolve lidar com valores ausentes, outliers e garantir que os dados estejam limpos e prontos para o modelo. Alguns aspectos importantes incluem:

-Imputação de Dados Ausentes:

Escolher estratégias adequadas para preencher valores ausentes, como a média, mediana ou métodos mais avançados.

-Detecção e Lidar com Outliers:

Identificar e tratar outliers, que podem afetar negativamente a performance do modelo.

Técnicas de Processamento:

-Além das estratégias mencionadas, existem técnicas mais avançadas de processamento de dados, como:

Redução de Dimensionalidade:

-***Utilizar técnicas como PCA (Análise de Componentes Principais)*** *para reduzir a dimensionalidade dos dados, mantendo a informação essencial.*

-Agrupamento de Variáveis:

Agrupar variáveis relacionadas para simplificar a representação do modelo.

Codificação de Variáveis Categóricas:

-Converter variáveis categóricas em formatos que o modelo possa entender, como one-hot encoding.

9. Passo a Passo da Exploração dos dados

- ***Compreensão dos Dados:***

data.head(): Exibe as primeiras linhas do conjunto de dados para entender a estrutura.
data.info(): Fornece informações sobre tipos de dados e valores ausentes.

- ***Limpeza dos Dados:***

    Lide com valores ausentes (NaN).
    Remova duplicatas.
    Converta tipos de dados.

data.dropna(): Remove linhas com valores ausentes. *Existem várias outras estratégias para tratar valores ausentes, dependendo do contexto.*

- ***Análise Descritiva:***

data.describe(): Apresenta estatísticas descritivas para variáveis numéricas, como média, desvio padrão, mínimo e máximo.

- ***Identificação de Padrões:***

Visualização da distribuição de uma variável usando um histograma.

- ***Análise de Anomalias:***

Detecção de outliers usando um boxplot.

- ***Análise de Relações:***

Visualização da relação entre duas variáveis usando um scatter plot.

10. Conceitos de correlação

Em uma análise complexa, é normal trabalharmos com um dataset que possua algumas dezenas de campos (colunas) e milhares de amostras(linhas). Algumas colunas do dataset serão relacionadas com outras, afinal foram coletadas do mesmo evento. Um desses campos do registro pode afetar ou não o valor de outro campo, ou seja, quando um campo aumenta de valor, acaba influenciando o valor do outro, seja para aumentar ou diminuir o seu quantitativo nominal.

A força do relacionamento entre duas colunas em um dataset é chamado de correlação, representada por um valor numérico entre -1 e 1.

É a medida de relacionamento mútuo entre duas variáveis, sejam causais ou não. Correlação ***pode existir entre quaisquer tipos de dados (contínuos ou categóricos). Apesar de representar um relacionamento mútuo, correlação não é sinônimo de causalidade.***

Pode ser útil em alguns casos, como o exemplo clássico do Sorvete. O produto é mais vendido em dias quentes. Ou seja, quanto maior a temperatura, maior o número de vendas de sorvete. Adicionalmente, se a relação entre essas variáveis é forte o suficiente, então nós podemos realizar previsões de comportamentos futuros.

Por exemplo, altura e peso são ambos relacionados; isto é, pessoas altas tendem a ser mais pesadas que pessoas baixas. Se surgir uma nova pessoa que é mais alta que a média das pessoas observadas até agora, então é bem provável que ela pese mais que a média das pessoas já observadas.

Correlação nos diz como as variáveis mudam em conjunto, ambas para a mesma direção ou em direções opostas, e a magnitude dessa relação. *Antes de mostrarmos o cálculo da correlação, precisamos entender o cálculo da covariância.*

Em estatística, a ***covariância é a medida de associação entre uma variável X e Y.*** *Para ser exato, mede a relação de tendência linear entre as variáveis, sendo calculada da seguinte forma:*

$$Cov(X, Y) = E[(X - E[X])(Y - E[Y])]$$

É calculado subtraindo cada item da variável pela sua média (centralizando os dados). Faz-se o produto entre esses dois valores centralizados. Por fim, calcula-se o valor esperado(E), essa variável é calculada em termos da média (μ).

Coeficiente de Pearson

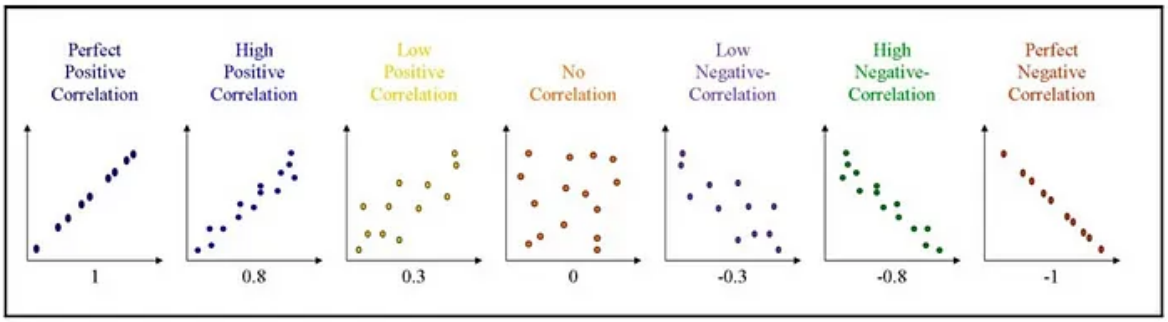
O coeficiente de pearson é um dos mais usados para cálculo da correlação. É a medida linear entre X e Y e varia entre -1 e 1. *É calculado dividindo-se a covariância pelo produto do desvio padrão de X(σX) e Y(σY).*

$$\rho_{X,Y} = \frac{E[(X-E[X])(Y-E[Y])]}{\sigma_X \sigma_Y}$$

A divisão pelo desvio padrão faz com que os dados fiquem entre o intervalo -1 e 1. Isso permite a comparação entre diversas variáveis, pois toda correlação ficará no mesmo intervalo.

Quando o coeficiente é próximo de 1, significa que há uma correlação positiva entre as variáveis X e Y. Uma correlação positiva indica que quando uma variável aumenta, a outra também aumenta. Por outro lado, quando mais perto de -1, a correlação é negativa. Ou seja, se uma variável aumenta, a outra diminui, e vice versa. Dizemos que X e Y são independentes quando o coeficiente de correlação é próximo de 0.

Utilizando scatter plots, podemos demonstrar alguns casos de como uma variavel é afetada por outra:



Perfect Positive Correlation:

Definição: Refere-se a uma relação linear exata positiva entre duas variáveis. Isso significa que, à medida que uma variável aumenta, a outra também aumenta na mesma proporção.

High (Alta) Positive Correlation:

Definição: Indica uma relação positiva forte entre duas variáveis, mas não necessariamente perfeita. À medida que uma variável aumenta, a outra tende a aumentar, mas pode haver alguma variabilidade.

Low Positive Correlation:

Definição: Refere-se a uma relação positiva fraca entre duas variáveis. O aumento em uma variável está associado a um aumento na outra, mas a relação não é muito forte.

No Correlation:

Definição: Indica a ausência de qualquer relação linear entre as duas variáveis. As mudanças em uma variável não estão associadas a mudanças sistemáticas na outra.

Low Negative Correlation:

Definição: Refere-se a uma relação negativa fraca entre duas variáveis. À medida que uma variável aumenta, a outra tende a diminuir, mas a relação não é muito forte.

High Negative Correlation:

Definição: Indica uma relação negativa forte entre duas variáveis, mas não necessariamente perfeita. À medida que uma variável aumenta, a outra tende a diminuir, mas pode haver alguma variabilidade.

Perfect Negative Correlation:

Definição: Refere-se a uma relação linear exata negativa entre duas variáveis. Isso significa que, à medida que uma variável aumenta, a outra diminui na mesma proporção.

11. Outliers

Outliers são valores extremos que diferem da maioria dos outros pontos de dados em um conjunto de dados. *Eles podem ter um grande impacto nas suas análises estatísticas e distorcer os resultados de quaisquer testes de hipóteses .*

É importante identificar cuidadosamente possíveis valores discrepantes em seu conjunto de dados e lidar com eles de maneira adequada para obter resultados precisos.

- ***Maneiras de calcular outliers***
  - **Ordenando valores quantitativos** (Você pode classificar variáveis quantitativas de forma crescente e procurar valores extremamente baixos ou extremamente altos. Sinalize quaisquer valores extremos que você encontrar.)

Seu conjunto de dados para um experimento piloto consiste em 8 valores.

180	156	9	176	163	1827	166	171

Você classifica os valores de baixo para alto e procura valores extremos.

9	156	163	166	171	176	180	**1872**

  - **Usando o intervalo interquartil para encontrar valores discrepantes**
    - **Classifique seus dados do menor para o maior.**

22	24	25	28	29	31	35	37	41	53	64

Organize seus dados em ordem crescente.

- **Identifique o primeiro quartil (Q1), a mediana e o terceiro quartil (Q3).**

Q1 é o valor abaixo do qual 25% dos dados estão localizados.

A mediana é o valor que divide os dados ao meio.

Q3 é o valor abaixo do qual 75% dos dados estão localizados.

Como você tem 11 valores, a mediana é o 6º valor. O valor mediano é 31.

22	24	25	28	29	**31**	35	37	41	53	64

A seguir, usaremos o método exclusivo para identificar Q1 e Q3. Isso significa que removemos a mediana dos nossos cálculos.

O Q1 é o valor no meio da primeira metade do seu conjunto de dados, excluindo a mediana. O valor do primeiro quartil é 25.

22	24	**25**	28	29

Seu valor do terceiro trimestre está no meio da segunda metade do seu conjunto de dados, excluindo a mediana. O valor do terceiro quartil é 41.

35	37	**41**	53	64

- **Calcule seu IQR (Intervalo Interquartil) = Q3 – Q1.**

O IQR representa a dispersão dos dados dentro da região central.

Fórmula	Cálculo
AIQ = Q3 − Q1	Q1 = 26 Q3 = 41 AIQ = 41 − 26 = 15

- **Calcule seu limite superior = Q3 + (1,5 * IQR).**

  Este é o limite superior além do qual um ponto é considerado um valor discrepante.

Fórmula	Cálculo
Cerca superior = Q3 + (1,5 * IQR)	Cerca superior = 41 + (1,5 * 15) = 41 + 22,5 = 63,5

- **Calcule seu limite inferior = Q1 – (1,5 * IQR).**

  Este é o limite inferior além do qual um ponto é considerado um valor discrepante.
  Use seus limites para destacar quaisquer valores discrepantes, todos os valores que estão fora de suas cercas.

  Valores discrepantes são aqueles que caem fora dos limites estabelecidos pelas cercas superior e inferior.

Fórmula	Cálculo
Cerca inferior = Q1 – (1,5 * IQR)	Cerca inferior = 26 – (1,5 * IQR) = 26 – 22,5 = 3,5

Volte ao conjunto de dados classificado da Etapa 1 e destaque quaisquer valores que sejam maiores que a cerca superior ou menores que a cerca inferior. Esses são seus valores discrepantes.

- Cerca superior = 63,5
- Cerca inferior = 3,5

22	24	25	28	29	31	35	37	41	53	**64**

Em geral, você deve tentar aceitar valores discrepantes tanto quanto possível, a menos que esteja claro que eles representam erros ou dados incorretos.

12. Referências

https://www.scribbr.com/statistics/outliers/

The Elements of Data Analytic Style - Jeff Leek

https://www.inf.ufsc.br/~marcelo.menezes.reis/Caps1_e_2.pdf

https://github.com/diasctiago/dio/blob/main/An%C3%A1lise%20de%20dados%20com%20Python%20e%20Pandas/EDA_DIO.ipynb

https://blog.xpeducacao.com.br/analise-de-dados-python/

https://blog.ploomes.com/analise-de-dados/

Python Para Análise de Dados: Tratamento de Dados com Pandas, NumPy & Jupyter - Wes Mckinney

https://ealexbarros.medium.com/introdu%C3%A7%C3%A3o-%C3%A0-correla%C3%A7%C3%A3o-589bdf8b2040#:~:text=%C3%89%20a%20medida%20de%20relacionamento,o%20exemplo%20cl%C3%A1ssico%20do%20Sorvete.