# O que é uma análise exploratória dos dados.

# Por quê fazer uma análise exploratória

# Quais as ferramentas utilizadas, tanto estatísticas quanto linguagens de programação

# Crie um passo a passo, em python utilizando uma base de dados da internet.

# Faça um call to action para as minhas redes sociais

# Coloque 5 hashtags que façam sentidos

Comporte-se com um escritor de artigos, que é um data scientist sênior e escreva o Artigo atendendo as regras abaixo

{REGRAS}

>No máximo 5 linhas por bloco de explicação

> Me explique de uma maneira informal, como se eu fosse um estudando iniciante em estatística e data Science

> O Bloco 3, podem ter até 10 linhas.

> O passo a passo não tem limite de linhas

> Os blocos que serão criados estão abaixo:

1 O que é uma análise exploratória dos dados.

2 Por quê fazer uma análise exploratória

3 Quais as ferramentas utilizadas, tanto estatísticas quanto linguagens de programação

4 Crie um passo a passo, em python utilizando uma base de dados da internet.

5 Faça um call to action para as minhas redes sociais

6 Coloque 5 hashtags que façam sentidos

Descreva melhor somente o código em python criado no bloco 4 anterior e mostre o porque foi feito cada coisa.

Crie uma conclusão para esse artigo por favor

Transformando dados brutos em insights valiosos.   
O papel da análise exploratória de dados.

Ilustração gerada pela lexica.art

Mesa com livros em cima

Descrição gerada automaticamente com confiança baixa

A análise exploratória de dados (ou EDA, do inglês Exploratory Data Analysis) é a primeira etapa de um projeto de Data Science. É quando a gente "brinca" com os dados para entender o que temos em mãos. Pensamos nisso como uma investigação preliminar, onde usamos gráficos, estatísticas descritivas e outras técnicas para identificar padrões, detectar anomalias, testar hipóteses e verificar suposições. Em resumo, é conhecer o terreno antes de começar a construção!

(Após mais de 3 anos no ramo da construção civil, as pessoas começam a falar assim.)

**Por que fazer uma análise exploratória?**

Fazer uma EDA é crucial porque nos ajuda a entender melhor os dados com os quais estamos trabalhando. Imagine tentar resolver um quebra-cabeça sem olhar para a imagem na caixa – complicado, né? A EDA nos dá uma visão clara dos dados, ajuda a identificar problemas como valores ausentes ou inconsistências, e nos orienta sobre quais transformações ou tratamentos serão necessários antes de aplicar modelos mais complexos.

**Tá, mas como eu faço isso?**

Bem, muitos estatísticos usaram papel, caneta e calculadoras para criar uma série de ferramentas úteis para tirar valor de um conjunto de dados, mas felizmente estamos em 2024. Hoje é usado uma combinação de ferramentas estatísticas e linguagens de programação. As ferramentas estatísticas incluem medidas como média, mediana, desvio padrão e gráficos como histogramas, boxplots e scatter plots. As linguagens de programação mais populares para EDA são Python e R. Em Python, bibliotecas como Pandas, NumPy, Matplotlib e Seaborn são extremamente úteis. Já no R, temos pacotes como dplyr, ggplot2 e tidyr, que facilitam muito o trabalho.

**Mas, se aprende mesmo fazendo.**

Então vamos fazer uma EDA usando Python e a famosa base de dados Titanic, disponível no Kaggle.

Texto

Descrição gerada automaticamente

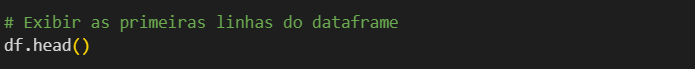
Para iniciar, é bom chamar a trinca mágica das bibliotecas para Ciência de Dados, dificilmente quem trabalha na área e lida com Python, não se deparou com elas em algum momento:

 Pandas: Para manipulação e análise de dados.

 NumPy: Para operações matemáticas e array.

 Matplotlib e Seaborn: Para visualização de dados.

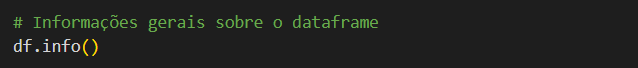
**df.head()**: Mostra as primeiras 5 linhas do DataFrame para termos uma visão inicial dos dados e das colunas presentes. Dessa forma, e junto com a descrição do dataset, da uma maior clareza daquilo que vamos trabalhar.



Tela de computador

Descrição gerada automaticamente

**df.info()**: Exibe um resumo informativo sobre o DataFrame, incluindo o número de entradas, nomes das colunas, tipos de dados e valores nulos. Isso ajuda a entender a estrutura dos dados.



Texto branco sobre fundo preto

Descrição gerada automaticamente

**df.describe():** Fornece estatísticas descritivas das colunas numéricas, como média, desvio padrão, mínimo, máximo e quartis. Isso nos dá uma visão geral da distribuição dos dados.



Tela de computador

Descrição gerada automaticamente

**df.isnull().sum():** Conta o número de valores ausentes (nulos) em cada coluna. Identificar valores ausentes é crucial para decidir como tratá-los (remoção, imputação, etc.).



Uma imagem contendo Tabela

Descrição gerada automaticamente

Texto

Descrição gerada automaticamente com confiança médiaGráfico, Histograma

Descrição gerada automaticamente

**sns.histplot(df['Age'].dropna(), kde=True)**: Cria um histograma para a coluna 'Age', excluindo valores nulos, com uma curva de densidade (KDE). Isso nos ajuda a visualizar a distribuição das idades dos passageiros.

Texto

Descrição gerada automaticamente

Gráfico, Gráfico de barras

Descrição gerada automaticamente

**sns.countplot(data=df, x='Sex', hue='Survived')**: Cria um gráfico de barras que mostra a contagem de sobreviventes e não sobreviventes por sexo. Isso nos permite ver se há uma diferença significativa na taxa de sobrevivência entre homens e mulheres.

Texto

Descrição gerada automaticamente

Gráfico, Gráfico de caixa estreita

Descrição gerada automaticamente

**sns.boxplot(data=df, x='Pclass', y='Age')**: Cria um boxplot que mostra a distribuição da idade dos passageiros em cada classe. Isso nos ajuda a entender a variação da idade entre diferentes classes socioeconômicas.

**Conclusão**

A análise exploratória de dados (EDA) é um passo fundamental em qualquer projeto de data science. Através dela, obtemos uma compreensão inicial dos dados, identificamos padrões e anomalias, e tomamos decisões informadas sobre como tratar e transformar os dados para as etapas seguintes. No exemplo do Titanic, vimos como ferramentas como Pandas, NumPy, Matplotlib e Seaborn podem ser utilizadas para realizar essa análise de maneira eficaz e eficiente.

Realizando uma EDA detalhada, conseguimos visualizar a distribuição de idades, entender a diferença na taxa de sobrevivência entre homens e mulheres, e explorar como a idade varia entre diferentes classes. Esses insights não só nos ajudam a preparar os dados para a modelagem, mas também a formular hipóteses e perguntas que podem ser respondidas em análises mais profundas.

Se você está começando na área de data science, dominar a EDA é essencial. Essa habilidade permitirá que você transforme dados brutos em insights valiosos, facilitando a construção de modelos preditivos robustos e confiáveis.

Gostou do conteúdo? Quer aprender mais sobre data science e estatística de uma forma descomplicada? Siga minhas redes sociais para mais dicas, tutoriais e insights exclusivos! Não perca nenhuma atualização!

Artigo criado com ajuda de Inteligência Artificial.

#DataScience #Estatística #Python #AnáliseDeDados #EDATutorial