Análise, construção e visualização de dados

Prof. Dr. Álvaro Campos Ferreira

alvaro.ferreira@idp.edu.br

Obtenção e Organização dos Dados

Tratamento e limpeza de dados

Qualidade de dados

Dimensões da qualidade de dados:

- Completude ou integridade
- Conformidade
- Validade
- Acurácia e precisão

Qualidade de dados

Soluções para quando já não se pode corrigir a coleta:

- Descartar o registro ruim
- Atribuir um valor de sentinela
- Atribuir o valor médio ou mais frequente
- Calcular um valor substituto
- Atribuir um valor baseado em valores vizinhos

idp

Valores faltantes

Valores faltantes são representados por **nan** em DataFrames e não são a mesma coisa que valores nulos.

- Faltas (completamente) aleatórias
- Faltas não aleatórias

Outliers

Outliers são valores encontrados fora de um intervalo razoável com relação às outras observações. É um ponto diferente dos demais, não necessariamente um erro. Sua identificação é importante e pode possuir informações interessantes sobre o sistema.

Visualização de dados

Visualização de dados

Podemos representar os dados de várias formas. Representações gráficas em geral facilitam a compreensão de suas características.

A visualização de dados deve ser de fácil atendimento, para que até mesmo as pessoas mais leigas no assunto possam entender a mensagem transmitida.

Visualização de dados

É preciso entender o que deve ser demonstrado e o público-alvo que se deseja atingir.

- Qual é o público-alvo?
- Que perguntas o gráfico deve responder?
- Que resposta o gráfico deve mostrar?
- Que mensagem deseja-se transmitir?

idp

Visualização de dados

Existem várias formas de se visualizar dados. Em geral, esses tipos se dividem em:

- Relacionamentos
- Comparação
- Distribuição
- Composição

Relacionamentos

Relacionamentos são os gráficos que mostram as relações entre dois conjuntos de dados.

- Scatter Plot
- Bubble Plot

Comparação

Para comparar dois conjuntos de dados graficamente, utiliza-se:

- Gráfico de linha (Line Plot)
- Gráfico de barras (Bar Plot)

Distribuição

Para visualizar a distribuição de valores,

- Histograma
- Box Plot

Propriedades

No Pandas, as características do gráfico gerado são controladas por suas propriedades.

- figsize
- title
- kind
- kind
- subplots
- grid
- xerr
- logx
- loglog
- legend
- xlim, ylim
- xlabel, ylabel

idp

Visualização de dados e inteligência de negócios (BI)

Business Intelligence (BI)

A inteligência de negócios (BI) é uma metodologia de coleta, análise e interpretação de dados relacionados ao problema de negócio de um empreendimento.

Geralmente se utiliza muito de visualização de dados, especialmente utilizando “dashboards”

INSTITUTO BRASILEIRO DE ENSINO,
DESENVOLVIMENTO E PESQUISA