Senac

Fecomércio Sesc

Data Science – Princípios e Técnicas

Setembro

2024

Data Science

Onde me encontrar:

https://www.linkedin.com/in/marco-mialaret-junior/

e

https://github.com/MatmJr

Voltando para o conjunto de dados das Faculdades Americanas

Data Science

```
import pandas as pd

url = 'http://personal.tcu.edu/kylewalker/data/colleges.csv'
data = pd.read_csv(url, encoding = 'latin_1')
data.shape
```

Data Science

Para responder a essa pergunta, precisamos identificar algumas colunas essenciais e filtrar os dados de acordo. As colunas que manteremos são as seguintes:

- **INSTNM**: Nome da instituição;

- **STABBR**: Estado onde a instituição está localizada;

Data Science

- **PREDDEG**: Tipo principal de diploma concedido pela instituição (códigos: 1 para certificados, 2 para diplomas de associado, 3 para bacharelados e 4 para pós-graduação);

- **CONTROL**: Propriedade da instituição (códigos: 1 para pública sem fins lucrativos, 2 para privada sem fins lucrativos, e 3 para privada com fins lucrativos);

Data Science

- **UGDS**: Número de alunos de graduação matriculados na instituição;

- **UG25abv**: Percentual de alunos de graduação com 25 anos ou mais na instituição.

Um pouco de estatística

Data Science

Nos últimos anos, houve um crescimento exponencial no volume de dados gerados pela humanidade, o que gerou uma demanda crescente por profissionais capazes de extrair informações e tomar decisões fundamentadas com base nesses dados.

Data Science

Um aspecto crucial ao trabalhar com dados é a habilidade de descrevê-los, resumi-los e representá-los visualmente. A estatística descritiva é uma ferramenta essencial nesse processo, utilizando duas abordagens principais:

Data Science

- **A abordagem quantitativa**, que descreve e resume os dados numericamente.

- **A abordagem visual**, que ilustra os dados por meio de gráficos e visualizações.

Data Science

Na análise quantitativa, destacamos:

- A **tendência central** informa sobre os centros dos dados. Medidas úteis incluem a média, mediana e moda.

- A **variabilidade** informa sobre a dispersão dos dados. Medidas úteis incluem variância e desvio padrão.

Data Science

- A **correlação** (ou **variabilidade conjunta**) informa sobre a relação entre um par de variáveis em um conjunto de dados. Medidas úteis incluem a covariância e o coeficiente de correlação.

As medidas de tendência central

Data Science

Média:

A média aritmética, ou simplesmente média, de um conjunto de valores é a medida de centro encontrada somando todos os valores do conjunto e dividindo pelo número de valores. Assim:

$$Média = \frac{soma\ dos\ valores}{total\ de\ observações}$$

Data Science

Exemplo 1: Determine a média de estudantes matriculados nas instituições.

Data Science

Obs: Existem outras médias, porém cada uma delas é usada em situações específicas. A saber:

- **Média Ponderada**: Você deve usar uma média ponderada quando deseja atribuir mais importância a alguns números em um conjunto de dados do que a outros. Isso é útil em cenários onde um evento pode ter vários resultados positivos ou negativos, e a magnitude desses resultados varia.

Data Science

- **Média Harmônica:** A média harmônica é calculada como o número de valores dividido pela soma do inverso de cada valor. É apropriada quando os dados representam grandezas que são inversamente proporcionais, como taxas.

Data Science

- **Média Geométrica:** A média geométrica é calculada como a raiz N-ésima do produto de todos os valores, onde N é o número de valores. É útil quando os dados estão em uma escala multiplicativa, como em situações envolvendo crescimento ou taxa de variação entre diferentes unidades de medida.

Mediana

A **mediana** da amostra é o elemento central de um conjunto de dados ordenado (crescente ou decrescente). Se o número de elementos n do conjunto de dados for ímpar, então a mediana é o valor na posição do meio. Se n for par, então a mediana é a média aritmética dos dois valores no meio

Data Science

Exemplo 2: Encontre a mediana do número de estudantes matriculados.

Data Science

Importante: A principal diferença entre o comportamento da média e da mediana está relacionada aos valores extremos (outliers) do conjunto de dados. De uma maneira geral:

Data Science

- Se você adicionar um valor discrepante maior do que a média em um conjunto de dados, a média aumentará, mas o valor da mediana vai sofrer pouca influência.

- Se você remover um valor discrepante de um conjunto de dados, a média diminuirá, mas o valor da mediana vai sofrer pouca influência.

Data Science

Agora vamos dividir o conjunto original em subconjuntos com características específicas e observar o que acontece com a média. Esse processo, conhecido como **estratificação**, consiste em separar os dados em subgrupos (ou estratos) que compartilham características semelhantes.

Data Science

Ao analisar cada estrato individualmente, podemos identificar variações internas e entender como cada subgrupo contribui para o comportamento geral dos dados, resultando em análises mais detalhadas e representativas.

Data Science

Exemplo 3: A média é mantida quando analisamos a quantidade de estudantes matriculados por tipo de diploma oferecido?

Moda

A **moda** da amostra é o valor no conjunto de dados que ocorre com mais frequência. Se não houver um único valor desse tipo, o conjunto será multimodal, pois possui vários valores modais.

Data Science

Exemplo 4: Qual a moda do número de estudantes matriculados.

Medidas de Localização

O percentil p da amostra é o elemento no conjunto de dados tal que p% dos elementos no conjunto de dados são menores ou iguais a esse valor. Além disso, (100 − p)% dos elementos são maiores ou iguais a esse valor. Se houver dois desses elementos no conjunto de dados, o percentil p da amostra é a média aritmética deles.

Data Science

- O primeiro quartil Q1 é o percentil 25 da amostra. Ele divide aproximadamente 25% dos menores itens do restante do conjunto de dados.

- O segundo quartil Q2 é o percentil 50 da amostra, também conhecido como a mediana. Aproximadamente 25% dos itens situam-se entre o primeiro e o segundo quartis, e outros 25% entre o segundo e o terceiro quartil.

Data Science

- O terceiro quartil Q3 é o percentil 75 da amostra. Ele divide aproximadamente 25% dos maiores itens do restante do conjunto de dados.

Data Science

Exemplo 6: Determine os quartis dos números de estudantes matriculados.

```
from statistics import quantiles

quantiles(dfNotNull.ugds, n=4,
method='inclusive')
```

Dúvidas?

Marco Mialaret, MSc

Telefone:
81 98160 7018

E-mail:
marcomialaret@gmail.com