**Một số hàm quan trọng**

1. Biểu đồ heatmap

* Trong visualization, heatmap là biểu đồ sử dụng cường độ màu sắc để thể hiện độ lớn của giá trị
* Ứng dụng:
  + Hiển thị ma trận tương quan
  + Hiện thị dữ liệu dạng ma trận
  + Theo dõi sự phân bố dữ liệu
* Đặc điểm chính của heatmap:
  + Màu sắc: các giá trị lớn sẽ được làm nổi bật bằng các vùng màu có cường độ ánh sáng mạnh, các giá trị nhỏ được thể hiện bằng các mảng màu nhạt hơn
  + Thang màu: Có thể điều chỉnh dựa trên các giá trị của dữ liệu, giúp người xem dễ dàng phân biệt được mức độ khác nhau

1. Các phương pháp thống kê

* Hàm mean()
  + Là hàm để tính giá trị trung bình bằng cách tính tổng của tất cả các giá trị của dữ liệu trong mẫu chia cho kích thước mẫu
* Hàm median()
  + Là hàm để tính giá trị trung vị bằng cách sắp xếp dữ liệu và lấy giá trị ở giữa. Nếu số giá trị là một số chẳn thì median là trung bình của 2 giá trị ở giữa
* Hàm mode()
  + Là hàm để tìm giá trị xuất hiện nhiều nhất trong một tập hợp dữ liệu
* Ba hàm trên sử dụng trong quá trình tiền xử lý dữ liệu, phân tích dữ liệu và đánh giá mô hình
* Hàm corr()
  + Trong lý thuyết xác suất và thống kê, hệ số tương quan (Coefficient Correlation) cho biết độ mạnh của mối quan hệ tuyến tính giữa hai biến số ngẫu nhiên.
  + Một khái niệm quan trọng khác liên quan đến tính tương quan là hiệp phương sai (covariance). Nếu phương sai dùng để đo lường sự biến thiên của một biến ngẫu nhiên (hay dữ liệu trên một tập mẫu) thì hiệp phương sai đo lường sự biến thiên của hai biến ngẫu nhiên (hay dữ liệu trên hai tập mẫu cùng số cá thể). Công thức hiệp phương sai của hai biến (hay hai tập mẫu có cùng n cá thể) x, y:

A black and white math equation

Description automatically generated

* Công thức tính hệ số tương quan dựa trên hiệp phương sai như sau:

A mathematical equation with black text

Description automatically generated

Với sdx và sdy tương ứng là độ lệch chuẩn của x và y.

1. Hàm StandardScaler()

hàm StandardScaler() thuộc thư viện scikit-learn được sử dụng phổ biến để chuẩn hóa dữ liệu trước khi đưa vào các mô hình học máy. Khi dữ liệu có các đặc trưng (features) với độ lớn hoặc thang đo khác nhau, việc sử dụng StandardScaler() giúp biến đổi các đặc trưng sao cho chúng có:

* Trung bình bằng 0 (mean = 0)
* Độ lệch chuẩn bằng 1 (standard deviation = 1)