Data visualization : trực quan hoá dữ liệu

Chọn biểu đồ : biểu đạt nhiều thông tin

Chính trị : bầu cử : vùng nào thiên hướng bầu cử nhiều hơn để lôi kéo ng dân bầu cử, nhìn giới tính, blabla -> nhiều cái khai thác, 1 phần của data science

Thời tiết : thu thập dữ liệu từ 100 năm về trc và truyền dữ liệu -> thông tin

Tín hiệu tim : đo đc -> thông tin -> truyền tải : anh nghiên cứu songs não : tìm đc tính chất liên quan não bộ : biểu đồ : songs cao : buồn ngủ, nói dối, blabla … -> mảng này hay

Quy trình của data science -> thu thập dữ liệu và phân tích -> thông tin -> visualize

Có thể áp dụng vào analyse và communicate từ hình ảnh -> phân tích

Có nhiều kiểu, dungf trên nhiều loại data -> master data -> làm data science, lương cao

Data engineer : ko mạnh = nc ngoài : thiên về làm công việc về data , AI -> bài toán khó, cần chuyên sâu AI -> việc ít, chưa phổ biến vn -> phân tích data quan trọng : kte, chứng khoán, chính trị, dữ liệu từ qk, tâm lý ng dung -> có thể phân tích, làm tỉ lệ -> có lãi

Import matplotlib.pylot as plt ; dung báo kỹ thuật

Import seaborn as sns : phân tích về xstk

Import plotly.express as px : làm về giao diện

Biểu đồ : so sánh, phân phối, liên hệ, thành phần : tròn cột miền …

Introduction : dataset : đếm cái nào k có dữ liệu : các sensor đo đạc, cái nào hỏng thì ô bị trống -> báo thông qua số, thiếu info -> nội suy hoặc bỏ đi : tg tác nào hợp lý thì bỏ vào, ko hợp thì bỏ đi (tính toán khoa học) : tránh thiếu vài cái mà xoá cả hàng, dữ liệu quý giữ lại

Data visualization for etth : choose the right chart type : chọn biểu đồ đugns

1. Xác dịnh data loại gì ; số chữ thgian, liên hệ các biến , trực quan để làm gì, người muốn show ra, xđ liên quan info nhiều hơn (sao kê) -> dựa vào mốc thgian để phântichs, ko gian : dựa vào vị trí để phân tích : song não dựa vào cực đẻ phân tích : thời tiết : dựa vào vị trí để phân tích
2. Identify the relationship between variable : so sánh tương quan : biểu đồ thường dùng 2 đến nhiều điểm -> có thể dung cột hoặc thanh, phân phối : data phân tán như nào : histogram : biểu đồ tần suất : relationship : bubble or histogram, area : chart or column
3. Dung biểu đồ phù hợp
4. Identify the audience : chọn biểu đồ : trình bày với khách : chỉ biểu đồ đơn giản, khách công nghệ hiểu biết sâu : cần hiểu sâu và đánh giá tốt hơn . đi gọi vốn : bgk ko biết về cntt, mà thuyết trình cn nhiều -> ko hiểu, cần trực quan hoá hơn để hiểu và không bị bắt bẻ, lại bị đánh giá kém -> cách trình bày
5. Select the appropriate chart type : đôi khi phải chọn nhiều biểu đồ khác nhau : phân tích về các câu hỏi : nếu thời gian sát nhau mà dung biểu đồ cột thì ko có ý nghĩa -> biểu đồ cột thiên về so sánh giữa các tính chất trong bảng -> so sánh cột này với cột này thì trực quan hơn, còn nếu so sánh cùng tính chất thì nhiều điểm, thì cần so sánh min hoặc tổng -> cần dung biểu đồ đường -> tổng hợp nhiều data, sẽ hiệu quả hơn

ETTh for dataset : phân phối thì dung biểu đồ histogram

Hiển thị phần trăm lỗi : trực quan hoá ra tính chất : nếu theo thgian thì bd đường , histogram -> cái nào max min thì cột, nhưng nên dungf đường thì tốt hơn vì thgian sát nhau, bđ độ tương quan : điểm or bubble, có thể dung các thành phần trực quan hoá để chiếm bao nhiêu phần tram : mỗi loại như thế nào

Biểu đồ đường : quan sát nhiều biến : quan sát xu hướng -> cách để plot điểm : matplotlib library : trục x date , y là value: lên xuống

Seaborn , plotly : liên quan hệ thống : dung thư viện cho web : dashboard có tính tương tác nháy động theo thời gian thực

Biểu đồ đường : subplot : số hàng và cột, chiều rộng chiều cao, hoặc có thể dung biểu đồ donut( same biểu đồ tròn)

Correlation chart : tương quan giữa các biến : đồng biến và nghịch biến : đưa vào ML và DL : cần phân tích xem có nên xoá những cái feature hay không

Scatter : dữ liệu phân bố thế nào : điểm chấm chấm : độ tương quan : biến thiên tăng -> đồng biến, giảm -> nghịch biến, tìm ra quy luật, yêu cầu cao về mặt triển khai về thời gian nhanh -> độ chính xác ko cao nhưng fps tốt hơn thay vì chính xác mà châm -> ko detect đc gì

Pair plot : theo từng cặp : tương quan theo trục hoành và tung : phân bố với nhau

Kaggle : data , tải data, thi Ai, tham khảo thú vị

Nêu tương quan 3 biến ; dung bubble : kgian 3 chiều

KDE chart : xem đc phân phối về mật độ, phân bố thế nào, các loại khác nhau có thể tổng hợp lại, xem tập trung về phần nào hơn

Displot chart : giống histogram nhưng rời rạc : xem được khoảng từ 0 – 10k bao người, blabla

Pairplot : iris : phân bố độ tương quan : các màu khác nhau : dữ liệu nhiều chiều, có thể dung chiều để gộp lại : ko gian 4 chiều để phân tích : 1 ô thì có thể phân tích 1 loại hoa, có thể dung để tách các đường biên của từng cụm

3D visualization : phân tích dạng 3d, tạo ra mp để cắt hoặc ko gian để tạo ra -> tạo model để phân loại. 4D : từ 3 chiều bd dạng vector + dung để bd dạng 4 chiều (lý thuyết) thực tế ko bd dạng 4 chiều, có thể dung nhiều cái 3 chiều để bd độ tương quan

Other visualization chart : word : phần nào quan tâm : từ nào nhiều thì bd to

Tip for selecting right chart : over time : biểu đồ điểm

Phân bố : histogram