[DA_Project TIMA] Task 6_Trực quan hóa dữ liệu

1. **Xác định mục tiêu trực quan hóa**
   - Mục tiêu: **Khám phá sự phân phối** của các biến số quan trọng trong bộ dữ liệu như `SoTienDKVayBanDau` (Số tiền đăng ký vay ban đầu), `TienGiaiNgan` (Tiền giải ngân) và `Salary` (Thu nhập).
   - Mục tiêu: **Phân tích mối quan hệ** giữa các biến liên tục như `SoTienDKVayBanDau` và `TienGiaiNgan`, và tìm mối liên kết giữa chúng với các biến phân loại như `Gender`, `Trạng thái`.
   - Mục tiêu: **Khám phá sự khác biệt** giữa các nhóm phân loại như `HasBadDebt` (Có nợ xấu) và `HasLatePayment` (Có thanh toán trễ).
   - Mục tiêu: **Phát hiện các giá trị ngoại lai** trong các biến như `Salary`, `TienGiaiNgan` để kiểm tra tính chính xác của dữ liệu.
2. **Chọn loại biểu đồ phù hợp với dữ liệu**
   - **Histogram**: Dùng để phân tích phân phối của các biến số liên tục như `SoTienDKVayBanDau`, `TienGiaiNgan`, và `Salary`.
   - **Boxplot**: Phát hiện các giá trị ngoại lai trong các biến liên tục như `SoTienDKVayBanDau`, `TienGiaiNgan`, `Salary`.
   - **Countplot**: Dùng để phân tích tần suất của các biến phân loại như `Gender`, `Trạng thái`, `CityName`, `JobName`.
3. **Trực quan hóa các biến số riêng biệt (Histogram, Boxplot, Countplot)**
   - **Histogram**:
     - Vẽ biểu đồ **Histogram** cho `SoTienDKVayBanDau` để kiểm tra phân phối của số tiền vay ban đầu.
     - Vẽ biểu đồ **Histogram** cho `TienGiaiNgan` để phân tích phân phối của tiền giải ngân.
     - Vẽ biểu đồ **Histogram** cho `Salary` để hiểu phân phối thu nhập của khách hàng.
   - **Boxplot**:
     - Vẽ **Boxplot** cho `SoTienDKVayBanDau` để phát hiện các giá trị ngoại lai trong số tiền vay.
     - Vẽ **Boxplot** cho `TienGiaiNgan` để kiểm tra các giá trị ngoại lai trong tiền giải ngân.
     - Vẽ **Boxplot** cho `Salary` để phân tích sự phân tán và các giá trị ngoại lai trong thu nhập.
   - **Countplot**:
     - Vẽ **Countplot** cho `Gender` để phân tích sự phân bố nam/nữ trong bộ dữ liệu.
     - Vẽ **Countplot** cho `Trạng thái` để kiểm tra sự phân bố của các trạng thái vay (Ví dụ: đang trả, đã trả xong).
     - Vẽ **Countplot** cho `CityName` và `JobName` để hiểu phân bố của khách hàng theo thành phố và nghề nghiệp.
4. **Khám phá mối quan hệ giữa các biến (Scatter plot, Heatmap, Pairplot)**
   - **Scatter plot**:
     - Vẽ **Scatter plot** giữa `SoTienDKVayBanDau` và `TienGiaiNgan` để kiểm tra mối quan hệ giữa số tiền vay và số tiền giải ngân.
     - Vẽ **Scatter plot** giữa `Salary` và `SoTienDKVayBanDau` để phân tích sự tương quan giữa thu nhập và khoản vay.
   - **Heatmap**:
     - Vẽ **Heatmap** để kiểm tra mối quan hệ giữa các biến liên tục như `SoTienDKVayBanDau`, `TienGiaiNgan`, `Salary`.
     - Vẽ **Heatmap** để phân tích sự tương quan giữa các biến phân loại như `Gender`, `CityName`, và các yếu tố tài chính.
   - **Pairplot**:
     - Vẽ **Pairplot** giữa các biến như `SoTienDKVayBanDau`, `TienGiaiNgan`, `Salary`, và `NumberOfLoans` để khám phá các mối quan hệ giữa các biến.
5. **Trực quan hóa mối quan hệ giữa nhiều biến**

- **Bubble chart**:
  - Vẽ **Bubble chart** để kiểm tra mối quan hệ giữa `SoTienDKVayBanDau`, `TienGiaiNgan` và `Salary`, với kích thước bong bóng thể hiện `NumberOfLoans`.
- **Stacked bar chart**:
  - Vẽ **Stacked bar chart** để phân tích số lượng các khoản vay trong các nhóm phân loại như `Gender`, `HasBadDebt` và `HasLatePayment`.
- **Clustered bar chart**:
  - Vẽ **Clustered bar chart** để so sánh số lượng khoản vay giữa các nhóm `CityName` và `Gender`.

6. **Trực quan hóa các nhóm phân loại (Bar plot, Violin plot)**
   - **Bar plot**:
     - Vẽ **Bar plot** để so sánh `Salary` giữa các nhóm `JobName` (Nghề nghiệp).
     - Vẽ **Bar plot** để phân tích sự khác biệt của `Salary` giữa các nhóm phân loại như `Gender`.
   - **Violin plot**:
     - Vẽ **Violin plot** để so sánh phân phối của `Salary` giữa các nhóm `CityName`.
     - Vẽ **Violin plot** để so sánh phân phối `SoTienDKVayBanDau` giữa các nhóm `ProductCreditName` (Tên sản phẩm tín dụng).
7. **Tinh chỉnh trực quan hóa (Màu sắc, Kích thước, Nhãn)**
   - Tinh chỉnh màu sắc của các biểu đồ để dễ dàng phân biệt các nhóm phân loại (Ví dụ: dùng màu sắc khác nhau cho `Gender` trong biểu đồ `Bar plot`).
   - Điều chỉnh kích thước các điểm trong biểu đồ **Scatter plot** để thể hiện mức độ quan trọng của các quan sát.
   - Thêm nhãn cho các trục trong biểu đồ **Boxplot** và **Histogram** để dễ dàng đọc được các giá trị.
   - Thêm tiêu đề và chú thích vào các biểu đồ để giúp người dùng hiểu rõ hơn về thông tin.
8. **Trực quan hóa dữ liệu theo thời gian (Line plot, Area plot)**
   - **Line plot**:
     - Vẽ **Line plot** để theo dõi sự thay đổi của `SoTienDKVayBanDau` theo thời gian, ví dụ theo tháng hoặc năm.
     - Vẽ **Line plot** để theo dõi sự thay đổi của `TienGiaiNgan` theo thời gian, nhằm phát hiện xu hướng.
   - **Area plot**:
     - Vẽ **Area plot** để thể hiện sự thay đổi của các khoản vay (`SoTienDKVayBanDau`) theo các thời kỳ, giúp đánh giá sự thay đổi của số tiền đăng ký vay.
9. **Đánh giá và điều chỉnh biểu đồ**
   - Kiểm tra độ chính xác của các dữ liệu trước khi điều chỉnh biểu đồ, để đảm bảo tính hợp lệ và chính xác của các biểu đồ.
   - Điều chỉnh các trục biểu đồ để phù hợp với phạm vi của dữ liệu (Ví dụ: thay đổi phạm vi trục Y cho các biến có phân phối rộng như `Salary`).
   - Đánh giá màu sắc, kiểu chữ và độ tương phản của các biểu đồ để đảm bảo chúng dễ đọc và dễ hiểu.
   - Kiểm tra lại các biểu đồ sau khi thay đổi các tham số để đảm bảo tính rõ ràng và dễ dàng tiếp cận.

Quy trình trực quan hóa

Xác định mục tiêu trực quan hóa --> **Chọn loại biểu đồ phù hợp với dữ liệu** --> **Trực quan hóa các biến số riêng biệt** (Histogram, Boxplot, Countplot) --> **Khám phá mối quan hệ giữa các biến** (Scatter plot, Heatmap, Pairplot) --> **Trực quan hóa mối quan hệ giữa nhiều biến** --> **Trực quan hóa các nhóm phân loại** (Bar plot, Violin plot) --> **Tinh chỉnh trực quan**

hóa (màu sắc, kích thước, nhãn) --> **Trực quan hóa dữ liệu theo thời gian** (Line plot, Area plot) --> **Đánh giá và điều chỉnh biểu đồ**

Quy trình chi tiết

1. Xác định mục tiêu trực quan hóa

- **Mục tiêu có thể là**:
  - Hiểu sự phân phối và phân tán của các biến số trong bộ dữ liệu.
  - Khám phá mối quan hệ giữa các biến liên tục hoặc phân loại.
  - Tìm kiếm sự khác biệt giữa các nhóm phân loại (ví dụ: nhóm khách hàng vay có **bad debt** và không có **bad debt**).
  - Phát hiện các giá trị ngoại lai hoặc bất thường trong dữ liệu.

2. Chọn loại biểu đồ phù hợp với dữ liệu

- **Dữ liệu liên tục** (ví dụ: **SoTienDKVayBanDau**, **TienGiaiNgan**, **Salary**):
  - **Histogram** và **KDE plot** để khám phá phân phối của các biến số liên tục.
  - **Boxplot** để phát hiện các giá trị ngoại lai (outliers).
- **Dữ liệu phân loại** (ví dụ: **Gender**, **Trạng thái**, **CityName**):
  - **Bar plot** hoặc **Count plot** để kiểm tra sự phân bổ tần suất của các nhóm phân loại.
- **Mối quan hệ giữa các biến số** (ví dụ: **SoTienDKVayBanDau** và **TienGiaiNgan**):
  - **Scatter plot** hoặc **Pairplot** để kiểm tra mối quan hệ giữa các biến số.

3. Trực quan hóa các biến số riêng biệt

- **Histogram** cho các biến liên tục như **SoTienDKVayBanDau**, **TienGiaiNgan** để xem phân phối và độ lệch của chúng.
- **Boxplot** cho các biến số liên tục để kiểm tra các ngoại lai và sự phân phối của chúng (ví dụ: **TienGiaiNgan**).
- **Count plot** cho các biến phân loại như **Trạng thái**, **Gender**, **CityName** để xem sự phân bố của các nhóm phân loại.

4. Khám phá mối quan hệ giữa các biến

- **Scatter plot**:
  - Để kiểm tra mối quan hệ giữa hai biến số liên tục, ví dụ: **SoTienDKVayBanDau** với **TienGiaiNgan**.
- **Heatmap**:
  - Dùng để kiểm tra mối quan hệ tương quan giữa các biến số liên tục (ví dụ: mối tương quan giữa **Salary**, **NumberOfLoans**, và **SoTienConLai**).
- **Pairplot**:
  - Nếu bạn muốn kiểm tra mối quan hệ giữa nhiều biến số liên tục, **Pairplot** sẽ giúp bạn so sánh tất cả các cặp biến liên tục trong dữ liệu (ví dụ: **SoTienDKVayBanDau**, **TienGiaiNgan**, **Salary**).

5. Trực quan hóa mối quan hệ giữa nhiều biến

- **Heatmap**:
  - Dùng để thể hiện ma trận tương quan giữa các biến số, ví dụ: **SoTienDKVayBanDau**, **TienGiaiNgan**, **Salary**.
- **Pairplot**:
  - Cũng có thể dùng để kiểm tra mối quan hệ giữa nhiều biến số liên tục như **SoTienDKVayBanDau**, **TienGiaiNgan**, **Salary** và **SoTienConLai**.

6. Trực quan hóa các nhóm phân loại

- **Bar plot**:
  - Dùng để trực quan hóa sự phân bổ của các nhóm phân loại, ví dụ: phân tích số lượng khách hàng vay theo các **Trạng thái** (đã trả nợ, chưa trả nợ).
  - Phân tích **Gender** hoặc **CityName** để thấy sự phân bố giữa các nhóm.
- **Violin plot**:
  - Dùng để so sánh sự phân phối của các biến số liên tục giữa các nhóm phân loại (ví dụ: so sánh **Salary** theo các nhóm **Gender**).

7. Tinh chỉnh trực quan hóa

- **Màu sắc**: Điều chỉnh màu sắc để phân biệt các nhóm trong biểu đồ (ví dụ: dùng màu khác nhau cho các nhóm **Trạng thái**, **Gender**).
- **Kích thước**: Thay đổi kích thước các điểm trong **Scatter plot** hoặc **Size plot** để phản ánh tầm quan trọng của các biến (ví dụ: **Salary**).
- **Nhãn**: Đảm bảo mỗi trục đều có nhãn rõ ràng và biểu đồ có tiêu đề mô tả chính xác nội dung.

8. Trực quan hóa dữ liệu theo thời gian

- **Line plot**:
  - Nếu dữ liệu có yếu tố thời gian, ví dụ như **CheckTime** hoặc các thông tin liên quan đến **Thời gian đã sống**, bạn có thể dùng **Line plot** để theo dõi sự thay đổi theo thời gian.
- **Area plot**:
  - Thích hợp cho việc thể hiện sự thay đổi trong các nhóm hoặc phân phối theo thời gian, ví dụ: sự thay đổi của **TienGiaiNgan** qua các năm.

9. Đánh giá và điều chỉnh biểu đồ

- **Kiểm tra rõ ràng và dễ hiểu**: Đảm bảo biểu đồ truyền tải đúng thông tin cần thiết và dễ hiểu đối với người xem.
- **Đánh giá tính chính xác của các biểu đồ**: Xem lại dữ liệu đã được trực quan hóa có đầy đủ và chính xác không, các mối quan hệ có được thể hiện rõ ràng không.

9.1 Biểu Đồ Cho Biến Số Liên Tục

Biểu đồ cho các biến số liên tục (số học, có thể chia nhỏ, như chiều cao, cân nặng, thu nhập,...).

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Histogram	Hiển thị phân phối của một biến số liên tục.	Biến liên tục	Nhận diện phân phối của dữ liệu, kiểm tra sự phân phố chuẩn hoặc lệch.
Boxplot	Hiển thị các yếu tố thống kê cơ bản như median, quartiles, outliers.	Biến liên tục	Phát hiện giá trị ngoại lai và đánh giá sự phân tán của dữ liệu.
Violin Plot	Kết hợp boxplot và KDE để thể hiện phân phối của dữ liệu.	Biến liên tục	Xem xét sự phân phối chi tiết hơn về mật độ xác suất của biến số liên tục.
Line Plot	Hiển thị dữ liệu theo chuỗi thời gian.	Biến liên tục theo thời gian	Dùng để theo dõi sự thay đổi của một biến theo thời gian (ví dụ: giá cổ phiếu).
Area Plot	Biểu đồ diện tích, thường được sử dụng với dữ liệu theo thời gian.	Biến liên tục theo thời gian	Tương tự line plot, nhưng có thể hiển thị sự tích lũy hoặc tổng hợp dữ liệu theo thời gian.

9.2 Biểu Đồ Cho Biến Số Phân Loại

Biểu đồ cho các biến phân loại (như giới tính, khu vực, loại sản phẩm,...).

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Countplot	Hiển thị số lượng mẫu của mỗi nhóm phân loại.	Biến phân loại	Kiểm tra số lượng các quan sát trong mỗi nhóm phân loại (ví dụ: số lượng khách hàng theo giới tính).
Bar Plot	Hiển thị giá trị trung bình, tổng hay bất kỳ thống kê nào của các nhóm phân loại.	Biến phân loại	So sánh các nhóm phân loại (ví dụ: so sánh thu nhập trung bình giữa các nhóm).
Pie Chart	Thể hiện tỷ lệ phần trăm của các nhóm phân loại trong một tập dữ liệu.	Biến phân loại	Hiển thị tỷ lệ phần trăm các nhóm trong một biến phân loại (ví dụ tỷ lệ giới tính).
Stacked Bar	Biểu đồ cột chồng, giúp phân tích sự phân bố của các nhóm phân loại theo các phân nhóm con.	Biến phân loại	Hiển thị tỷ lệ của các phân nhóm trong từng nhóm phân loại.
Heatmap	Hiển thị ma trận sự tương quan giữa các nhóm phân loại hoặc biến số.	Biến phân loại	Phân tích mối quan hệ giữa các nhóm phân loại hoặc giữa các nhóm phân loại và biến liên tục.

9.3 Biểu Đồ Cho Mối Quan Hệ Giữa Các Biến

Biểu đồ này dùng để khám phá mối quan hệ giữa các biến số (liên tục hoặc phân loại).

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Scatter Plot	Hiển thị mối quan hệ giữa hai biến liên tục.	Biến liên tục	Kiểm tra sự tương quan giữa hai biến (ví dụ: chiều cao và cân nặng).
Pairplot	Hiển thị mối quan hệ giữa tất cả các cặp biến trong một bộ dữ liệu.	Biến liên tục hoặc phân loại	Phân tích mối quan hệ giữa nhiều biến, tìm sự tương quan giữa các biến số khác nhau.
Heatmap	Hiển thị ma trận tương quan giữa các biến.	Biến liên tục	Phát hiện mối tương quan giữa nhiều biến số và kiểm tra sự đồng biến của chúng.
Bubble Chart	Biểu đồ phân tán với kích thước vòng tròn thay đổi theo giá trị của một biến thứ ba.	Biến liên tục	Khám phá mối quan hệ giữa ba biến, thông qua tọa đ (x, y) và kích thước (kích thước của vòng tròn).

9.4 Biểu Đồ Cho Dữ Liệu Theo Thời Gian

Dữ liệu chuỗi thời gian thường được thể hiện qua các biểu đồ thích hợp để theo dõi sự thay đổi của một hoặc nhiều biến theo thời gian.

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Line Plot	Hiển thị sự thay đổi của biến theo thời gian.	Biến liên tục theo thời gian	Phân tích xu hướng theo thời gian, ví dụ như giá trị cổ phiếu t tháng.
Area Plot	Hiển thị diện tích dưới đường line plot, thể hiện sự thay đổi theo thời gian.	Biến liên tục theo thời gian	Hiển thị sự thay đổi và tích lũy theo thời gian, ví dụ doanh thu trong các tháng.
Stacked Area Plot	Biểu đồ diện tích chồng, cho phép phân tích nhiều nhóm cùng lúc theo thời gian.	Biến liên tục theo thời gian	So sánh sự thay đổi của nhiều nhóm dữ liệu theo thời gian (ví doanh thu của các sản phẩm khác nhau).

9.5 Biểu Đồ Cho Mối Quan Hệ Giữa Các Nhóm

Dùng để phân tích các mối quan hệ giữa các nhóm phân loại, đặc biệt khi cần so sánh các đặc điểm của từng nhóm.

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Violin Plot	Biểu đồ violin, kết hợp giữa boxplot và KDE để thể hiện phân phối của một biến số cho các nhóm phân loại.	Biến liên tục, phân loại	So sánh phân phối của một biến liên tục trong các nhóm phân loại, dễ dàng nhận diệ khác biệt.
Bar Plot	Biểu đồ cột để so sánh các giá trị của nhóm phân loại.	Biến phân loại	So sánh các giá trị tổng hợp hoặc trung bình của các nhó phân loại.

9.6 Biểu Đồ Cho Dữ Liệu Phân Tán (Categorical vs Continuous)

Để phân tích mối quan hệ giữa một biến phân loại và một biến liên tục.

Biểu đồ	Chức năng	Kiểu Dữ Liệu	Mục Đích Sử Dụng
Boxplot	Hiển thị mối quan hệ giữa một biến phân loại và biến liên tục.	Biến phân loại và liên tục	So sánh phân phối của biến liên tục trong c nhóm phân loại.
Violin Plot	Tương tự như boxplot nhưng hiển thị phân phối mật độ của biến liên tục cho các nhóm phân loại.	Biến phân loại và liên tục	Đánh giá phân phối và sự dày đặc của các nhóm phân loại trên biến liên tục.