[DA_Project TIMA] Task 7_Phân tích Mô tả

Phân tích Mô tả

1. Tính giá trị trung bình của "Số tiền vay"

- **Mục đích:** Để xác định số tiền vay trung bình của khách hàng.
- **Hàm cần sử dụng:** mean()
- **Ý nghĩa của hàm:** mean() tính giá trị trung bình của cột "Số tiền vay".

2. Tính độ lệch chuẩn của "Lãi suất"

- **Mục đích:** Để phân tích mức độ phân tán của lãi suất.
- **Hàm cần sử dụng:** std()
- **Ý nghĩa của hàm:** std() tính độ lệch chuẩn của cột "Lãi suất", cho thấy mức độ phân tán của giá trị.

3. Tính tổng "Tiền giải ngân"

- **Mục đích:** Xác định tổng số tiền đã giải ngân.
- **Hàm cần sử dụng:** sum()
- **Ý nghĩa của hàm:** sum() tính tổng giá trị của cột "Tiền giải ngân".

4. Hiển thị các thống kê mô tả cho "Điểm tín dụng"

a. **Mục đích:** Để có cái nhìn tổng quan về phân phối điểm tín dụng của khách hàng.
b. **Hàm cần sử dụng:** describe()
c. **Ý nghĩa của hàm:** describe() trả về các thống kê mô tả bao gồm giá trị trung bình, độ lệch chuẩn, min, max, và các phần trăm phân vị của "Điểm tín dụng".

5. Tính số lượng khách hàng theo từng "Trạng thái"

a. **Mục đích:** Xác định số lượng khách hàng trong mỗi trạng thái.
b. **Hàm cần sử dụng:** value_counts()
c. **Ý nghĩa của hàm:** value_counts() đếm số lần xuất hiện của mỗi trạng thái trong cột "Trạng thái".

6. Vẽ biểu đồ phân phối cho "Thu nhập"

a. **Mục đích:** Để trực quan hóa phân phối thu nhập của khách hàng.
b. **Hàm cần sử dụng:** hist()
c. **Ý nghĩa của hàm:** hist() vẽ biểu đồ histogram để trực quan hóa phân phối của cột "Thu nhập".

7. Vẽ biểu đồ boxplot cho "Tiền giải ngân"

a. **Mục đích:** Phân tích sự phân bố của số tiền giải ngân.
b. **Hàm cần sử dụng:** boxplot()
c. **Ý nghĩa của hàm:** boxplot() vẽ biểu đồ hộp giúp phân tích sự phân tán và các giá trị ngoại lệ trong "Tiền giải ngân".

8. Vẽ biểu đồ hình tròn cho tỷ lệ "Giới tính"

a. **Mục đích:** Để phân tích tỷ lệ khách hàng theo giới tính.

b. **Hàm cần sử dụng:** pie()

c. **Ý nghĩa của hàm:** pie() tạo ra biểu đồ hình tròn để trực quan hóa tỷ lệ các giá trị trong cột "Giới tính".

9. Tính giá trị trung bình của "Lãi suất" theo từng "Ngành nghề"

a. **Mục đích:** Để phân tích mức lãi suất trung bình theo ngành nghề.

b. **Hàm cần sử dụng:** groupby(), mean()

c. **Ý nghĩa của hàm:** groupby() nhóm dữ liệu theo ngành nghề và mean() tính giá trị trung bình của lãi suất trong mỗi nhóm ngành nghề.

10. Tính số lượng khách hàng theo "Khu vực"

- **Mục đích:** Để xác định số lượng khách hàng theo khu vực.
- **Hàm cần sử dụng:** groupby(), size()
- **Ý nghĩa của hàm:** groupby() nhóm dữ liệu theo khu vực và size() tính số lượng khách hàng trong mỗi khu vực.

11. Vẽ biểu đồ phân tán giữa "Lương" và "Điểm tín dụng"

- **Mục đích:** Để kiểm tra mối quan hệ giữa lương và điểm tín dụng của khách hàng.
- **Hàm cần sử dụng:** scatter(), sns.scatterplot()
- **Ý nghĩa của hàm:** scatter() vẽ biểu đồ phân tán giúp tìm hiểu mối quan hệ giữa lương và điểm tín dụng.

12. Tính số lượng "Nợ xấu"

- **Mục đích:** Để đếm số lượng khách hàng có nợ xấu.
- **Hàm cần sử dụng:** value_counts()
- **Ý nghĩa của hàm:** value_counts() đếm số lượng các giá trị trong cột "Nợ xấu", giúp phân tích tỷ lệ nợ xấu.

13. Tính trung bình "Thời gian đã sống" theo "Ngành nghề"

- **Mục đích:** Xác định mức độ trung bình về thời gian đã sống của khách hàng trong mỗi ngành nghề.
- **Hàm cần sử dụng:** groupby(), mean()
- **Ý nghĩa của hàm:** groupby() nhóm dữ liệu theo ngành nghề và mean() tính giá trị trung bình của thời gian đã sống.

14. Vẽ biểu đồ heatmap cho "Lãi suất" và "Điểm tín dụng"

- **Mục đích:** Để phân tích mối quan hệ giữa lãi suất và điểm tín dụng.
- **Hàm cần sử dụng:** sns.heatmap()
- **Ý nghĩa của hàm:** sns.heatmap() giúp vẽ biểu đồ nhiệt (heatmap) để thấy mối quan hệ giữa các biến.

15. Tính tổng "Lãi suất" theo "Khu vực"

- **Mục đích:** Để tổng hợp lãi suất theo từng khu vực.
- **Hàm cần sử dụng**: groupby(), sum()
- **Ý nghĩa của hàm:** groupby() nhóm dữ liệu theo khu vực và sum() tính tổng lãi suất trong mỗi nhóm khu vực.

16. Tính số lượng "Khách hàng theo ngành nghề"

- **Mục đích:** Để phân tích số lượng khách hàng trong mỗi ngành nghề.
- **Hàm cần sử dụng:** groupby(), size()
- **Ý nghĩa của hàm:** groupby() nhóm theo ngành nghề và size() tính số lượng khách hàng trong mỗi ngành nghề.

17. Hiển thị các giá trị lớn nhất của "Số tiền vay"

- **Mục đích:** Để phân tích các giá trị lớn nhất trong số tiền vay.
- **Hàm cần sử dụng:** nlargest()
- **Ý nghĩa của hàm**: nlargest() trả về các giá trị lớn nhất từ cột "Số tiền vay".

18. Hiển thị các giá trị nhỏ nhất của "Lãi suất"

- **Mục đích:** Để phân tích các giá trị nhỏ nhất của lãi suất.
- **Hàm cần sử dụng:** nsmallest()
- **Ý nghĩa của hàm:** nsmallest() trả về các giá trị nhỏ nhất từ cột "Lãi suất".

19. Kiểm tra sự phân phối của "Điểm tín dụng"

- **Mục đích:** Để xác định sự phân phối của điểm tín dụng.
- **Hàm cần sử dụng**: sns.distplot()
- **Ý nghĩa của hàm:** sns.distplot() tạo biểu đồ phân phối giúp phân tích sự phân bố của điểm tín dụng.

20. Tính số lượng "Số lượng khoản vay" theo "Thành phố"

- **Mục đích:** Để phân tích số lượng các khoản vay theo từng thành phố.
- **Hàm cần sử dụng:** groupby(), size()
- **Ý nghĩa của hàm:** groupby() nhóm theo thành phố và size() tính số lượng các khoản vay trong từng thành phố.

21. Vẽ biểu đồ phân phối thu nhập của khách hàng theo nhóm tuổi

- **Mục đích:** Để phân tích phân phối thu nhập theo nhóm tuổi.
- **Hàm cần sử dụng**: sns.boxplot()
- **Ý nghĩa của hàm:** sns.boxplot() vẽ biểu đồ hộp giúp so sánh phân phối thu nhập trong các nhóm tuổi.

22. Vẽ biểu đồ tần suất "Giới tính"

- **Mục đích:** Để phân tích tỷ lệ giới tính trong dữ liệu.
- **Hàm cần sử dụng:** value_counts(), plot()
- **Ý nghĩa của hàm:** value_counts() đếm số lượng các giá trị và plot() vẽ biểu đồ tần suất.

23. Kiểm tra sự tương quan giữa "Điểm tín dụng" và "Số tiền vay"

- **Mục đích:** Xem xét sự tương quan giữa điểm tín dụng và số tiền vay.
- **Hàm cần sử dụng:** corr()
- **Ý nghĩa của hàm:** corr() tính toán ma trận tương quan giữa các cột trong dữ liệu.

24. Phân tích số liệu theo loại "Hình thức cư trú"

- **Mục đích:** Để phân tích các đặc điểm của khách hàng theo hình thức cư trú.
- **Hàm cần sử dụng:** groupby(), size()
- **Ý nghĩa của hàm:** groupby() nhóm theo "Hình thức cư trú" và size() tính số lượng khách hàng trong mỗi nhóm.

25. Tính tổng số "Khoản vay" theo từng "Phường"

- **Mục đích:** Để phân tích số lượng khoản vay tại từng phường.
- **Hàm cần sử dụng:** groupby(), sum()

- **Ý nghĩa của hàm:** groupby() nhóm theo phường và sum() tính tổng số khoản vay trong mỗi phường.

Tóm tắt luồng phân tích

Xác định vấn đề → Thu thập dữ liệu → Làm sạch dữ liệu → Tính toán thống kê cơ bản → Chọn loại biểu đồ → Trực quan hóa dữ liệu → Phân tích và đánh giá kết quả → Kiểm tra giả thuyết và mối quan hệ nhân quả → Cung cấp giải pháp hành động.

Quy trình chi tiết

1. **Xác định vấn đề**
   - **Mục tiêu**: Phân tích mối quan hệ giữa các yếu tố tài chính và hành vi vay vốn của khách hàng.
     - **Vấn đề cụ thể**: Xác định xem liệu có mối quan hệ giữa **số tiền vay ban đầu** (SoTienDKVayBanDau) và **điểm tín dụng** (TS_CREDIT_SCORE_V2), cũng như ảnh hưởng của **mức thu nhập** (Salary) và **địa phương cư trú** (CityName, DistrictName) đến quyết định vay vốn của khách hàng.
2. **Thu thập dữ liệu**
   - **Nguồn dữ liệu**: Bộ dữ liệu liên quan đến thông tin khách hàng, các khoản vay, số tiền vay ban đầu, thu nhập, điểm tín dụng, và các yếu tố khác như khu vực cư trú, nghề nghiệp, v.v.
     - Ví dụ: Các bảng hoặc tập tin dữ liệu có thể bao gồm các cột như:
       - SoTienDKVayBanDau: Số tiền vay ban đầu của khách hàng.
       - TS_CREDIT_SCORE_V2: Điểm tín dụng của khách hàng.
       - Salary: Mức thu nhập của khách hàng.
       - CityName: Thành phố nơi khách hàng sinh sống.
       - DistrictName: Quận/huyện nơi khách hàng sinh sống.
3. **Làm sạch dữ liệu**
   - **Xử lý giá trị thiếu**: Kiểm tra các cột có dữ liệu thiếu như Salary, TS_CREDIT_SCORE_V2 và thay thế hoặc loại bỏ giá trị thiếu.
     - Ví dụ: Nếu Salary bị thiếu, có thể thay thế bằng giá trị trung bình hoặc loại bỏ các dòng chứa giá trị thiếu.
   - **Xử lý dữ liệu ngoại lai (outliers)**: Kiểm tra xem có các giá trị ngoài phạm vi hợp lý (ví dụ: điểm tín dụng vượt quá 900 hoặc số tiền vay bất thường).
     - Có thể sử dụng phương pháp **boxplot** để phát hiện các ngoại lai trong các cột số như SoTienDKVayBanDau.
   - **Chuẩn hóa dữ liệu**: Đảm bảo các cột dữ liệu có đúng kiểu dữ liệu, ví dụ: Salary phải là số thực, TS_CREDIT_SCORE_V2 là số nguyên.
4. **Tính toán thống kê cơ bản**
   - **Mục tiêu**: Hiểu rõ các đặc điểm cơ bản của dữ liệu thông qua các chỉ số thống kê.
     - **mean()**: Tính giá trị trung bình của các cột như Salary, SoTienDKVayBanDau, TS_CREDIT_SCORE_V2.
     - **std()**: Tính độ lệch chuẩn để hiểu mức độ phân tán của các biến này.
     - **describe()**: Đưa ra bảng thống kê mô tả cho tất cả các cột số trong bộ dữ liệu (ví dụ: Salary, SoTienDKVayBanDau).
     - **median()**: Xác định giá trị trung vị của các cột số liệu, ví dụ: TS_CREDIT_SCORE_V2 để hiểu phân phối dữ liệu.
     - **value_counts()**: Đếm tần suất xuất hiện của các giá trị trong các cột phân loại như CityName hoặc Gender.
5. **Chọn loại biểu đồ**
   - **Histogram**: Hiển thị phân phối của dữ liệu cho các biến số liên tục như Salary và SoTienDKVayBanDau.
     - Biểu đồ này giúp nhận diện độ lệch, độ phân tán và sự đồng nhất của dữ liệu.
   - **Boxplot**: Phân tích phân phối và phát hiện giá trị ngoại lai của TS_CREDIT_SCORE_V2 và Salary. Giúp phát hiện các điểm bất thường hoặc giá trị cực trị.

- **Scatter Plot**: Kiểm tra mối quan hệ giữa các cặp biến liên tục như TS_CREDIT_SCORE_V2 và SoTienDKVayBanDau. Biểu đồ này giúp xác định có mối quan hệ tuyến tính hoặc phi tuyến giữa hai biến.
- **Bar Chart**: So sánh các nhóm dữ liệu, ví dụ, sử dụng **Bar Chart** để so sánh số tiền vay ban đầu theo các thành phố hoặc quận (CityName, DistrictName).
- **Heatmap**: Hiển thị mối quan hệ tương quan giữa các biến số trong bộ dữ liệu, ví dụ: tương quan giữa Salary, TS_CREDIT_SCORE_V2, và SoTienDKVayBanDau.
- **Violin Plot**: Trực quan hóa sự phân bố của các cột số và phát hiện các giá trị ngoại lai. Ví dụ, so sánh sự phân bố của Salary cho các nhóm Gender hoặc CityName.

6. **Trực quan hóa dữ liệu**
   - **Dựng biểu đồ**:
     - **Histogram** cho Salary để xem phân phối thu nhập.
     - **Boxplot** cho TS_CREDIT_SCORE_V2 và SoTienDKVayBanDau để phát hiện ngoại lai.
     - **Scatter plot** cho mối quan hệ giữa TS_CREDIT_SCORE_V2 và SoTienDKVayBanDau.
     - **Heatmap** cho ma trận tương quan để xem mối quan hệ giữa các biến tài chính (ví dụ, Salary, TS_CREDIT_SCORE_V2 và SoTienDKVayBanDau).

- **Tạo báo cáo**: Các biểu đồ trực quan giúp thấy rõ các xu hướng và mối quan hệ trong dữ liệu.

7. **Phân tích và đánh giá kết quả**
   - **Nhận diện mối quan hệ**: Dựa trên các biểu đồ, đánh giá mối quan hệ giữa các biến. Ví dụ:
     - Liệu có mối quan hệ tuyến tính giữa điểm tín dụng và số tiền vay không? Biểu đồ **scatter plot** sẽ cho thấy nếu có mối quan hệ đó.
     - Sự phân phối của thu nhập và số tiền vay có bị lệch? Biểu đồ **boxplot** hoặc **violin plot** sẽ giúp kiểm tra sự đồng đều trong phân phối.
   - **Đánh giá ngoại lai**: Từ biểu đồ **boxplot**, xác định các giá trị ngoại lai, có thể là các khách hàng có điểm tín dụng hoặc số tiền vay bất thường.
8. **Kiểm tra giả thuyết và mối quan hệ nhân quả**
   - **Giả thuyết**: Xác định các giả thuyết cần kiểm tra. Ví dụ:
     - "Điểm tín dụng cao sẽ liên quan đến số tiền vay thấp."
     - "Thu nhập cao sẽ có ảnh hưởng đến số tiền vay ban đầu."
   - **Kiểm tra mối quan hệ nhân quả**: Sử dụng các biểu đồ như **sns.regplot()** hoặc **sns.lmplot()** để kiểm tra mối quan hệ nhân quả giữa các biến.
     - Nếu có mối quan hệ nhân quả, có thể thực hiện các kiểm tra thống kê như **Linear Regression** để xác nhận giả thuyết.
9. **Cung cấp giải pháp hành động**
   - **Giải pháp hành động**:
     - Nếu phát hiện rằng các khách hàng có điểm tín dụng thấp có xu hướng vay nhiều tiền hơn, một giải pháp có thể là cải thiện các chương trình đào tạo tài chính cho khách hàng này để nâng cao điểm tín dụng.
     - Nếu thu nhập có ảnh hưởng lớn đến số tiền vay, các chính sách cho vay có thể được điều chỉnh để phục vụ tốt hơn cho các nhóm thu nhập thấp

Danh sách hàm chính cần sử dụng:

1. Nhóm hàm tóm tắt thông tin thống kê cơ bản

Hàm: describe(), mean(), std(), sum(), median()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để tóm tắt các đặc điểm cơ bản của dữ liệu số trong DataFrame. Các hàm này giúp bạn nắm bắt các chỉ số thống kê cơ bản, cung cấp cái nhìn tổng quan về phân phối và sự phân tán của dữ liệu.
- **Chức năng**:
  - **describe()**: Trả về các chỉ số thống kê tổng quát như số lượng giá trị (count), giá trị trung bình (mean), độ lệch chuẩn (std), giá trị nhỏ nhất (min), các phân vị (25%, 50%, 75%), và giá trị lớn nhất (max).
  - **mean()**: Tính giá trị trung bình của một cột.
  - **std()**: Tính độ lệch chuẩn, đo mức độ phân tán của dữ liệu xung quanh giá trị trung bình.
  - **sum()**: Tính tổng các giá trị trong một cột.
  - **median()**: Tính giá trị trung vị, giá trị giữa của dữ liệu khi đã được sắp xếp theo thứ tự.

2. Nhóm hàm phân tích phân phối dữ liệu

Hàm: value_counts(), hist(), boxplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích phân phối của dữ liệu. Chúng giúp bạn hiểu sự phân bổ của các giá trị trong cột, phát hiện các giá trị ngoại lệ, và nhận diện các đặc điểm phân phối như độ lệch và sự phân bố đồng đều.
- **Chức năng**:
  - **value_counts()**: Đếm số lần xuất hiện của mỗi giá trị trong một cột. Thích hợp cho việc phân tích dữ liệu phân loại.
  - **hist()**: Vẽ biểu đồ histogram để hiển thị phân phối của một biến số. Biểu đồ này giúp nhận diện sự phân bố của dữ liệu, ví dụ: có lệch trái, lệch phải hay phân phối đều.
  - **boxplot()**: Vẽ biểu đồ hộp để phân tích sự phân bố và phát hiện các giá trị ngoại lệ. Biểu đồ này cung cấp cái nhìn rõ ràng về các phân vị và phạm vi của dữ liệu.

3. Nhóm hàm phân tích mối quan hệ giữa các biến

Hàm: corr(), sns.regplot(), sns.scatterplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích và xác định mối quan hệ giữa các biến trong dữ liệu. Chúng giúp phát hiện các mối quan hệ tuyến tính hoặc phi tuyến tính, từ đó xây dựng các mô hình dự báo hoặc tìm hiểu sự tương quan giữa các yếu tố.
- **Chức năng**:
  - **corr()**: Tính toán ma trận tương quan giữa các biến số. Hệ số tương quan cho biết mức độ mạnh yếu của mối quan hệ giữa các biến.
  - **sns.regplot()**: Vẽ biểu đồ hồi quy tuyến tính giữa hai biến, kết hợp với đường hồi quy. Giúp phân tích mối quan hệ tuyến tính giữa các biến.
  - **sns.scatterplot()**: Vẽ biểu đồ phân tán (scatter plot) giữa hai biến để nhìn nhận trực quan mối quan hệ giữa chúng.

4. Nhóm hàm nhóm và phân tích dữ liệu theo các nhóm

Hàm: groupby(), pivot_table(), sns.barplot()

- **Ý nghĩa**: Nhóm hàm này cho phép bạn nhóm dữ liệu theo các đặc tính nhất định và áp dụng các phép toán tổng hợp để phân tích sự khác biệt giữa các nhóm. Chúng rất hữu ích trong việc so sánh và phân tích các nhóm con của dữ liệu.
- **Chức năng**:
  - **groupby()**: Nhóm dữ liệu theo một hoặc nhiều cột và tính toán các giá trị thống kê cho mỗi nhóm, như trung bình, tổng, v.v.
  - **pivot_table()**: Tạo bảng tổng hợp (pivot table) từ dữ liệu, cho phép tính toán các chỉ số tổng hợp như trung bình, tổng, v.v., theo các nhóm và phân loại.

- **sns.barplot()**: Vẽ biểu đồ cột để so sánh giá trị trung bình của các nhóm dữ liệu, hỗ trợ so sánh giữa các nhóm một cách trực quan.

5. Nhóm hàm trực quan hóa dữ liệu

Hàm: sns.distplot(), sns.heatmap(), sns.pairplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để trực quan hóa dữ liệu dưới dạng đồ họa, giúp bạn dễ dàng hiểu và phân tích các mối quan hệ và phân phối trong dữ liệu.
- **Chức năng**:
  - **sns.distplot()**: Vẽ biểu đồ phân phối kết hợp với đường mật độ, giúp bạn hiểu rõ hơn về phân phối của dữ liệu.
  - **sns.heatmap()**: Vẽ heatmap, một biểu đồ thể hiện mối quan hệ giữa các biến thông qua bảng màu. Thường được sử dụng để trực quan hóa ma trận tương quan hoặc các bảng số liệu phức tạp.
  - **sns.pairplot()**: Vẽ biểu đồ phân tán của từng cặp biến trong DataFrame, giúp bạn dễ dàng phát hiện các mối quan hệ giữa các biến trong dữ liệu.

6. Nhóm hàm phân tích phân phối và phát hiện giá trị ngoại lệ

Hàm: sns.boxplot(), sns.violinplot()

- **Ý nghĩa**: Nhóm hàm này được sử dụng để phân tích sự phân bố của dữ liệu và phát hiện các giá trị ngoại lệ (outliers). Các biểu đồ này giúp bạn nhận diện những điểm bất thường trong dữ liệu và phân tích các phân phối một cách trực quan.
- **Chức năng**:
  - **sns.boxplot()**: Vẽ biểu đồ hộp (box plot) để phân tích sự phân bố của dữ liệu và phát hiện các giá trị ngoại lệ. Biểu đồ này cho thấy các phân vị (quartile) và các giá trị ngoài phạm vi bình thường.
  - **sns.violinplot()**: Vẽ biểu đồ violin để hiển thị sự phân phối của dữ liệu. Nó kết hợp giữa biểu đồ hộp và biểu đồ mật độ, giúp cung cấp cái nhìn sâu sắc hơn về phân phối của dữ liệu.