**REPORT OF FINAL EXAM OF DATA VISUALIZATION**

***Dataset: Default\_of\_credit\_card\_clients***

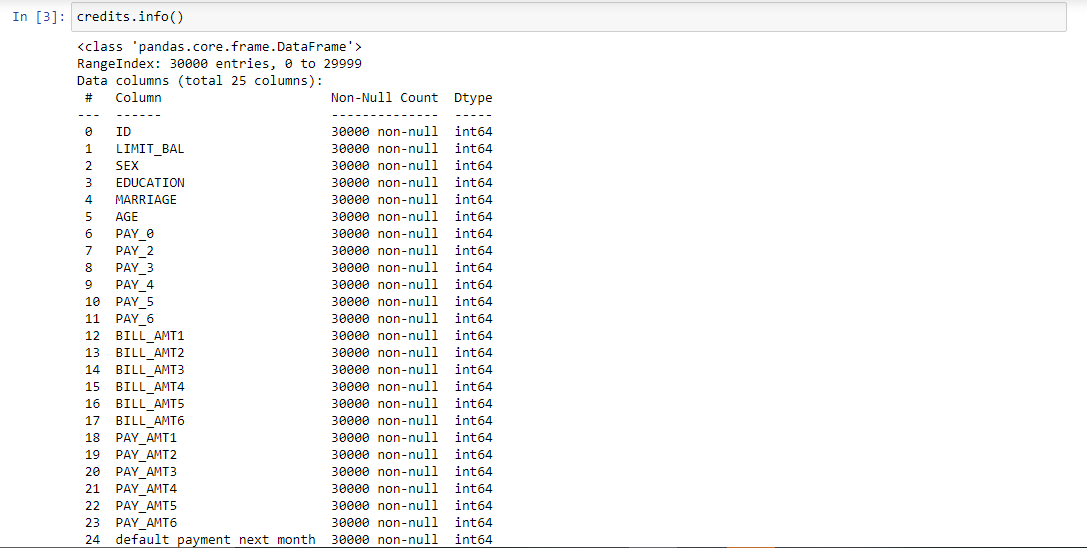
1. **Tìm hiểu dữ liệu:**

Chúng ta tìm hiểu định nghĩa của từng dữ liệu để dự đoán thanh toán tín dụng của khách hàng vào tháng 10 của dataset. Ta nhìn vào từng feature để mô tả cụ thể.

Tập dữ liệu có tổng cộng 25 cột, target là ‘***default.payment.next.month*’**

* ID: ID của từng khách hàng
* LIMIT\_BAL : Số tiền thẻ tín dụng được cấp đơn vị NT dollars (bao gồm tín dụng cá nhân và tín dụng gia đình/ tín dụng bổ sung
* SEX: Giới tích (1=nam, 2=nữ)
* EDUCATION : (1=cao học, 2=đại học, 3=cáp 3, 4=còn lại, 5=không xác định, 6= không xác định)
* MARRIAGE : Tình trạng hôn nhân(1=đã kết hôn, 2=độc thân, 3=khác)
* AGE : tuổi
* PAY\_0 : Tình trạng trả nợ tháng 9,2005 (-1=thanh tóa hợp lệ, 1=trì hoãn trả nợ 1 tháng, 2= trì hoãn trả nợ 2 tháng, ... 8= trì hoãn trả nợ 8 tháng, 9= trì hoãn trả nợ trên 9 tháng)
* PAY\_2 : Tình trạng trả nợ tháng 8,2005 (như trên)
* PAY\_3 : Tình trạng trả nợ tháng 7,2005 (như trên)
* PAY\_4 : Tình trạng trả nợ tháng 6,2005 (như trên)
* PAY\_5 : Tình trạng trả nợ tháng 5,2005 (như trên)
* PAY\_6 : Tình trạng trả nợ tháng 4,2005 (như trên)
* BILL\_AMT1 : Số tiền của sao kê hóa đơn tháng 9,2005 (NT dollar)
* BILL\_AMT2 : Số tiền của sao kê hóa đơn tháng 8,2005 (NT dollar)
* BILL\_AMT3 : Số tiền của sao kê hóa đơn tháng 7,2005 (NT dollar)
* BILL\_AMT4 : Số tiền của sao kê hóa đơn tháng 6,2005 (NT dollar)
* BILL\_AMT5 : Số tiền của sao kê hóa đơn tháng 5,2005 (NT dollar)
* BILL\_AMT6 : Số tiền của sao kê hóa đơn tháng 4,2005 (NT dollar)
* PAY\_AMT1 : Số tiền đã thanh toán trước đó vào tháng 9, 2005 (NT dollar)
* PAY\_AMT2 : Số tiền đã thanh toán trước đó vào tháng 8, 2005 (NT dollar)
* PAY\_AMT3 : Số tiền đã thanh toán trước đó vào tháng 7, 2005 (NT dollar)
* PAY\_AMT4 : Số tiền đã thanh toán trước đó vào tháng 6, 2005 (NT dollar)
* PAY\_AMT5 : Số tiền đã thanh toán trước đó vào tháng 5, 2005 (NT dollar)
* PAY\_AMT6 : Số tiền đã thanh toán trước đó vào tháng 4, 2005 (NT dollar)
* default.payment.next.month: Phương thức thanh toán mặc định(1=có, 0=không)

Đầu tiên import các thư viện để load dataset, tiếp theo dùng lệnh info() ta thấy dữ liệu đã ở dạng số (Numerical Data) nên không cần chuyển kiểu dữ liệu.

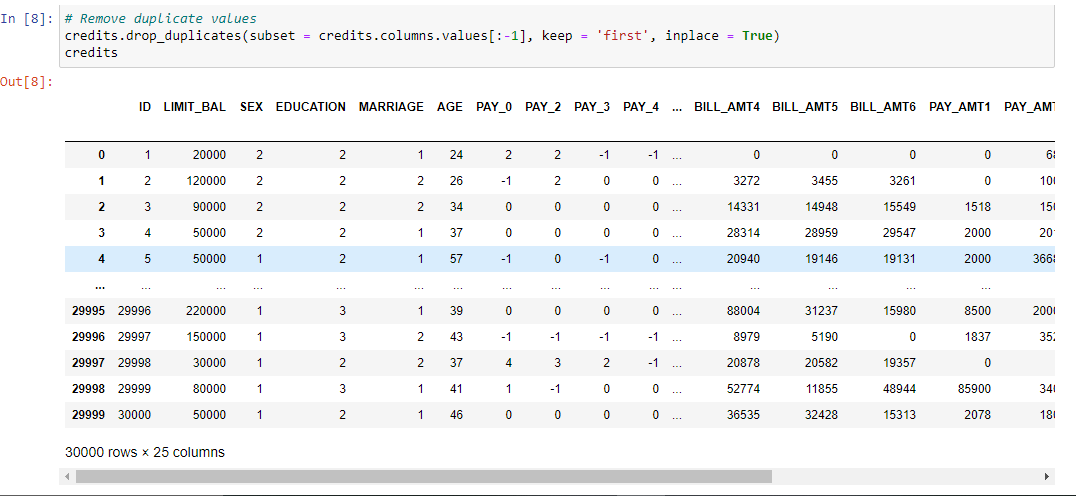


Sau đó ta sẽ đi tính missing values của dataset ở (out 5), dữ liệu không có các giá trị missing

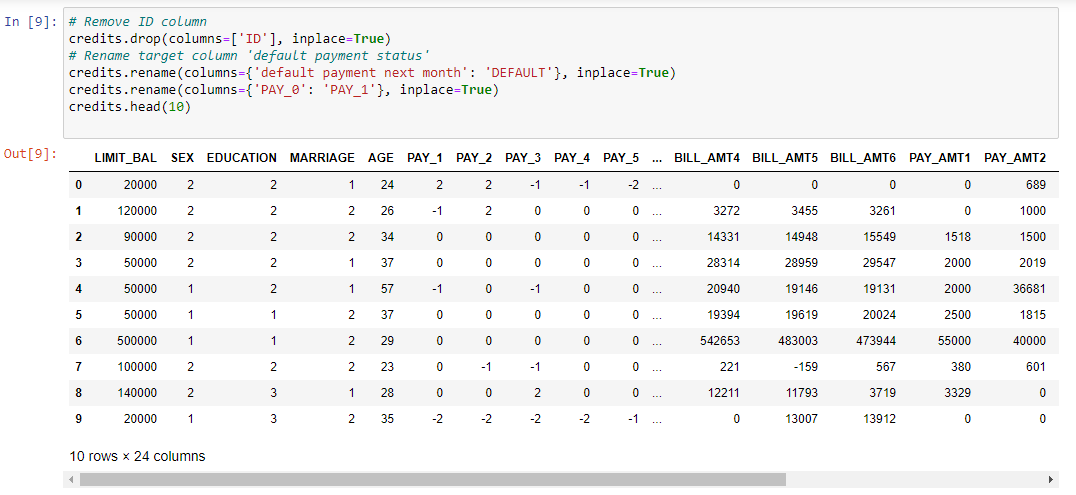
|  | **Missing values of credits dataset** | **Percentage Missing Values** |
| --- | --- | --- |
| **default payment next month** | 0 | 0.0 |
| **PAY\_6** | 0 | 0.0 |
| **LIMIT\_BAL** | 0 | 0.0 |
| **SEX** | 0 | 0.0 |
| **EDUCATION** | 0 | 0.0 |
| **MARRIAGE** | 0 | 0.0 |
| **AGE** | 0 | 0.0 |
| **PAY\_0** | 0 | 0.0 |
| **PAY\_2** | 0 | 0.0 |
| **PAY\_3** | 0 | 0.0 |
| **PAY\_4** | 0 | 0.0 |
| **PAY\_5** | 0 | 0.0 |
| **BILL\_AMT1** | 0 | 0.0 |
| **PAY\_AMT6** | 0 | 0.0 |
| **BILL\_AMT2** | 0 | 0.0 |
| **BILL\_AMT3** | 0 | 0.0 |
| **BILL\_AMT4** | 0 | 0.0 |
| **BILL\_AMT5** | 0 | 0.0 |
| **BILL\_AMT6** | 0 | 0.0 |
| **PAY\_AMT1** | 0 | 0.0 |
| **PAY\_AMT2** | 0 | 0.0 |
| **PAY\_AMT3** | 0 | 0.0 |
| **PAY\_AMT4** | 0 | 0.0 |
| **PAY\_AMT5** | 0 | 0.0 |
| **ID** | 0 | 0.0 |

1. **Làm sạch sữ liệu:**

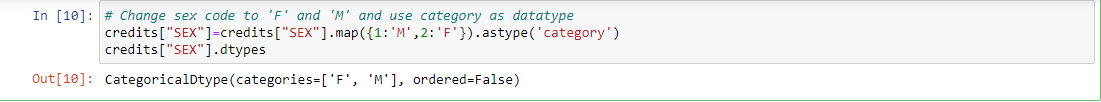
* Dùng lệnh “*credits.drop\_duplicates (subset = credits.columns.values[:-1], keep = 'first', inplace = True)”* để xóa bỏ các giá trị trùng lặp



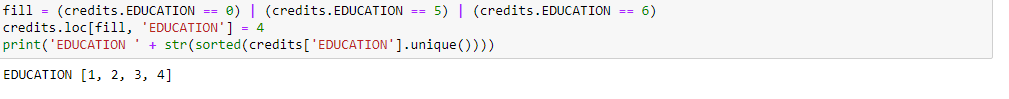
* Dùng lệnh credits.drop(columns=['ID'], inplace=True) để xóa bỏ cột ID. Ta xóa cột này vì dữ liệu từ ban đầu đã được xử lý và sắp xếp theo thứ tự nên ta có thể suy ra id từ index của nó.
* Target của dataset là cột *'default payment next month’*, ta đổi tên cột *'default payment next month'* thành *'DEFAULT’* cho gọn hơn. Sau đó đổi cột *'PAY\_0'* thành *'PAY\_1'*  để phù hợp với *'BILL\_AMT1'* and *'PAY\_AMT1'*.



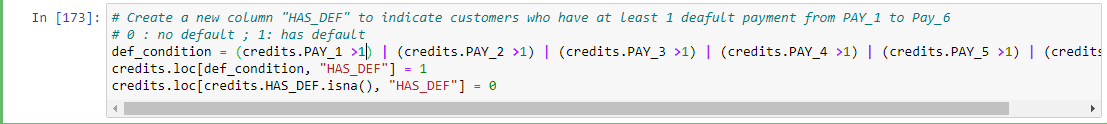
* Ở cột SEX: Giới tích (1=nam, 2=nữ) nên ta sẽ đổi *‘1’* thành *‘M’*(Male) *‘2’* là *‘F’*(Female)



* Ở cột EDUCATION: (1=cao học, 2=đại học, 3=cấp 3, 4=khác, 5=không xác định, 6= không xác định) ta sẽ thay các giá trị 0, 5, 6 quy về về4 để đại diện cho còn lại

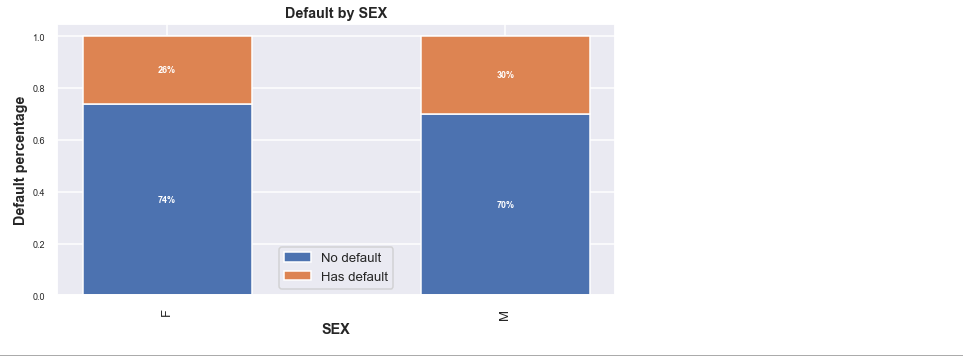
1. **Thống kê, phân tích và vẽ biểu đồ:**

* Đầu tiên ta sẽ tạo một cột mới "HAS\_DEF" để cho khách hang thấy được it nhất một thanh toán mặc định từ PAY\_1 to Pay\_6
* 0 : no default , 1: has default

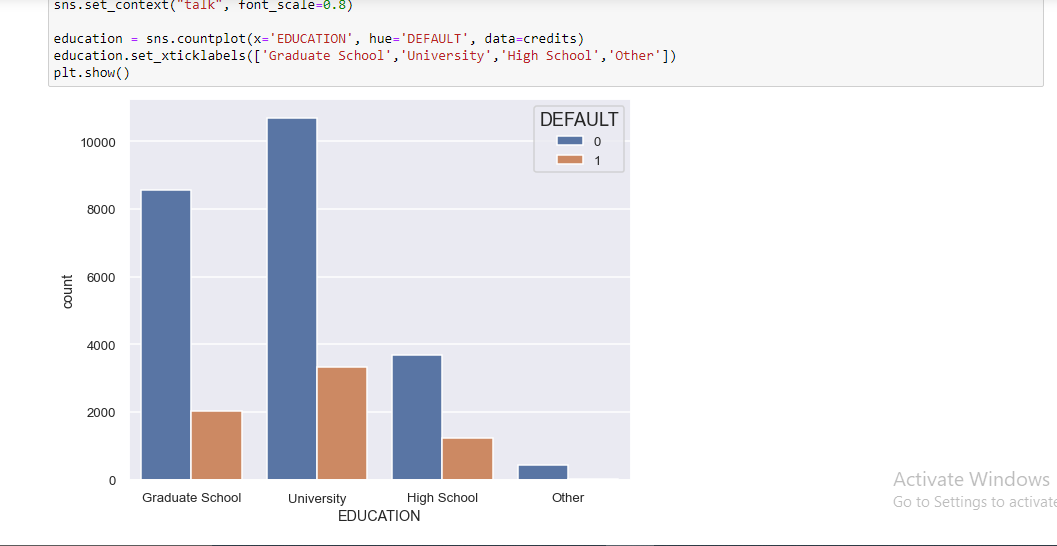


1. **Thống kê của cột SEX :**

* Bởi vì trong dữ liệu số lượng nữ nhiều số lượng nam nên ta sử dụng phần trăm cho mỗi nhóm giới tính. Ta thấy được 30% đàn ông có thanh toán mặc định trong khi đó 26% phụ nữ thanh toán mặc định. Sự chênh lệch không đáng kể.
* Trong dữ liệu này, đàn ông trễ thanh toán hơn phụ nữ. Điều này chỉ đúng cho tập dữ liệu này,không có ý nghĩa tương tự cho tập dữ liệu khác.



1. **Thống kê của cột EDUCATION:**



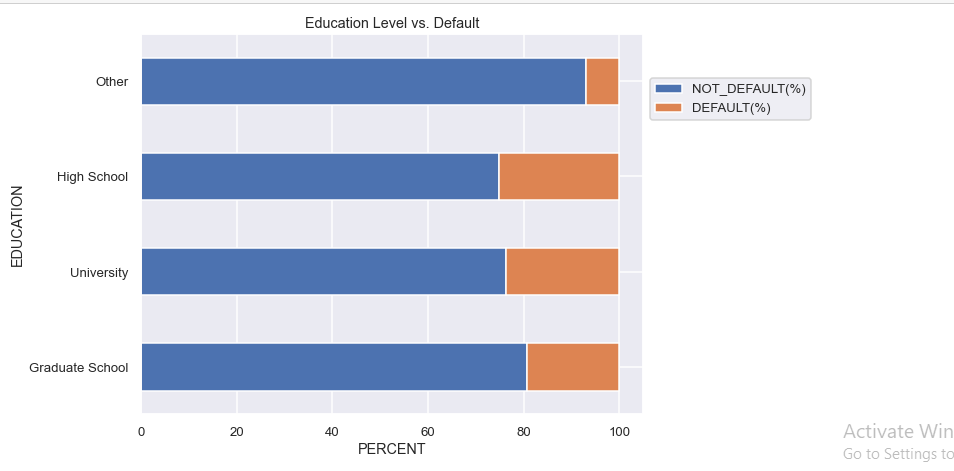
* Đầu tiên ta tính số lượng Default và Not Default của các bậc học Cao học, Đại học, Cấp 3 và còn lại.

|  | **EDUCATION** | **NOT\_DEFAULT** | **DEFAULT** | **TOTAL** |
| --- | --- | --- | --- | --- |
| **0** | Graduate School | 8549 | 2036 | 10585 |
| **1** | University | 10700 | 3330 | 14030 |
| **2** | High School | 3680 | 1237 | 4917 |
| **3** | Other | 435 | 33 | 468 |

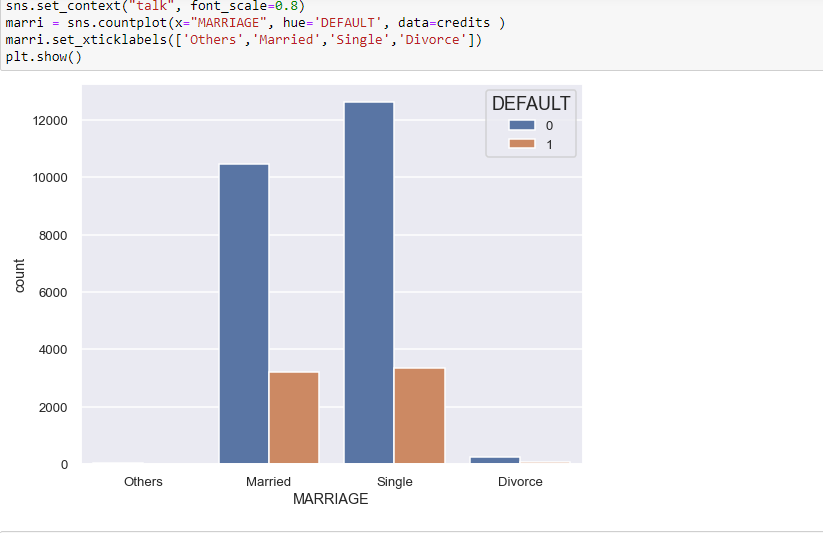
*Tiếp theo chuyển bảng trên theo lượng phần trăm*

|  | **EDUCATION** | **NOT\_DEFAULT(%)** | **DEFAULT(%)** |
| --- | --- | --- | --- |
| **0** | Graduate School | 80.77 | 19.23 |
| **1** | University | 76.27 | 23.73 |
| **2** | High School | 74.84 | 25.16 |
| **3** | Other | 92.95 | 7.05 |

* Sau đó ta dùng biểu đồ chồng để trực quan tỷ lệ phần trăm của dữ liệu.
* Mặc dù số lượng Defalult của High School thấp hơn Graduate School và University. Nhưng phần trăm Default của High School chiếm phần lớn, cao hơn Default của Graduate School và University.
* Những khách hàng có bậc học cao hơn ít dùng thanh toán mặc định và có hạn mức tìn dụng cao hơn.



1. **Thống kê của cột MARRIAGE:**

****

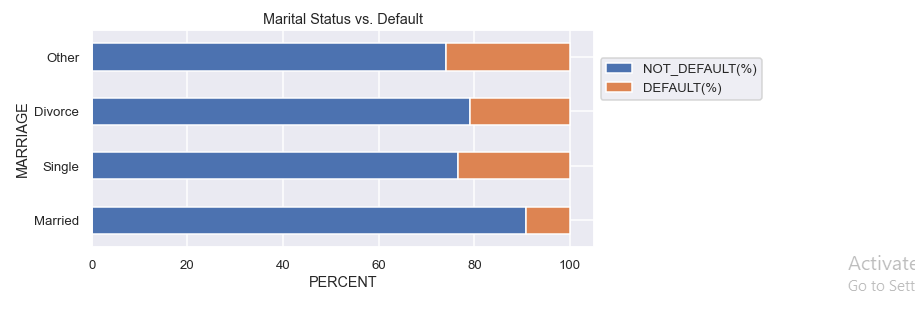
* Dường như không có mối tương quan giữa thanh toán mặc định và tình trạng hôn nhân.
* Như ở trên ta tính số lượng Default và Not Default của các Đã kết hôn, Độc thân, Ly hôn và còn lại. Tiếp theo chuyển bảng trên theo lượng phần trăm

|  | **MARRIAGE** | **NOT\_DEFAULT** | **DEFAULT** | **TOTAL** |
| --- | --- | --- | --- | --- |
| **0** | Married | 49 | 5 | 54 |
| **1** | Single | 10453 | 3206 | 13659 |
| **2** | Divorce | 12623 | 3341 | 15964 |
| **3** | Other | 239 | 84 | 323 |

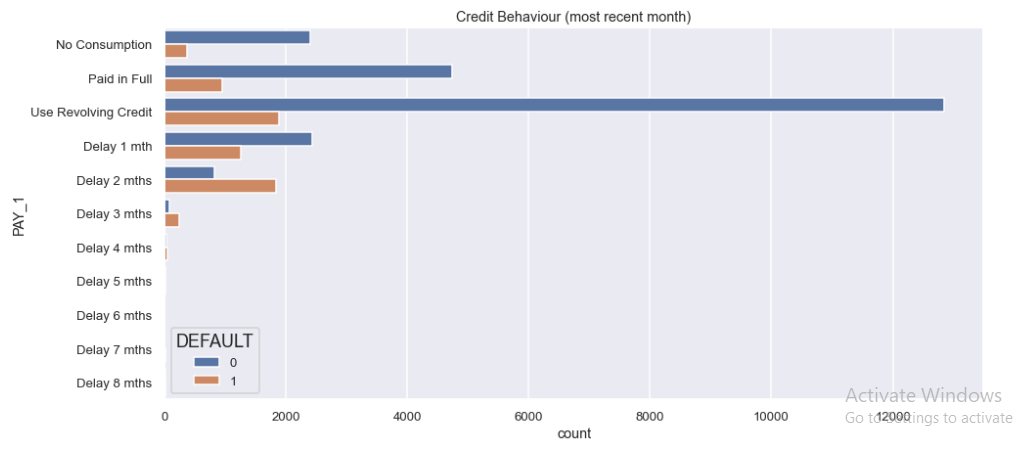
*Tiếp theo chuyển bảng trên theo lượng phần trăm*

| **MARRIAGE** | **NOT\_DEFAULT(%)** | **DEFAULT(%)** |
| --- | --- | --- |
| **0** | Married | 90.74 | 9.26 |
| **1** | Single | 76.53 | 23.47 |
| **2** | Divorce | 79.07 | 20.93 |
| **3** | Other | 73.99 | 26.01 |

* Dùng biểu đồ chồng để trực quan tỷ lệ phần trăm của dữ liệu.
* Mặc dù số lượng Default của Divorce thấp hơn Single & Married Nhưng phần trăm Default của Divorce chiếm cao hơn Default của Married.



1. **Thống kê của Hoạt Động Tín Dụng của khách hàng:**

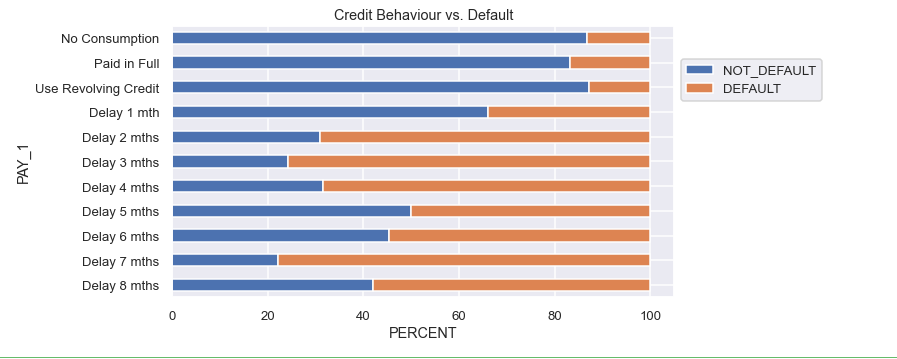
****

* Tính số lượng Default và Not Default của Không tiêu thụ, Thanh toán đủ, Sử dụng tín dụng xoay vòng, Chậm 1 tháng đến Chậm 8 tháng. Tiếp theo chuyển bảng trên theo lượng phần trăm

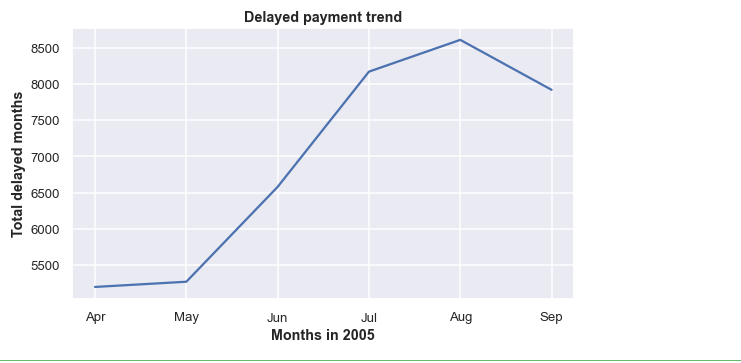
|  | **PAY\_1** | **NOT\_DEFAULT** | **DEFAULT** | **TOTAL** |
| --- | --- | --- | --- | --- |
| **0** | No Consumption | 2394 | 365 | 2759 |
| **1** | Paid in Full | 4732 | 954 | 5686 |
| **2** | Use Revolving Credit | 12849 | 1888 | 14737 |
| **3** | Delay 1 mth | 2436 | 1252 | 3688 |
| **4** | Delay 2 mths | 823 | 1844 | 2667 |
| **5** | Delay 3 mths | 78 | 244 | 322 |
| **6** | Delay 4 mths | 24 | 52 | 76 |
| **7** | Delay 5 mths | 13 | 13 | 26 |
| **8** | Delay 6 mths | 5 | 6 | 11 |
| **9** | Delay 7 mths | 2 | 7 | 9 |
| **10** | Delay 8 mths | 8 | 11 | 19 |

|  | **PAY\_1** | **NOT\_DEFAULT(%)** | **DEFAULT(%)** |
| --- | --- | --- | --- |
| **0** | No Consumption | 86.77 | 13.23 |
| **1** | Paid in Full | 83.22 | 16.78 |
| **2** | Use Revolving Credit | 87.19 | 12.81 |
| **3** | Delay 1 mth | 66.05 | 33.95 |
| **4** | Delay 2 mths | 30.86 | 69.14 |
| **5** | Delay 3 mths | 24.22 | 75.78 |
| **6** | Delay 4 mths | 31.58 | 68.42 |
| **7** | Delay 5 mths | 50.00 | 50.00 |
| **8** | Delay 6 mths | 45.45 | 54.55 |
| **9** | Delay 7 mths | 22.22 | 77.78 |
| **10** | Delay 8 mths | 42.11 | 57.89 |

* Dùng biểu đồ chồng để trực quan tỷ lệ phần trăm của dữ liệu.
* Những người sữ dụng Revolving Credit (chỉ trả tối thiểu) và những người bị trì hoãn trong 2 tháng thì có số lượng Default cao nhất. Khi thanh toán bị trì hoãn hơn 2 tháng, các cơ hội của Default cao hơn 50%.

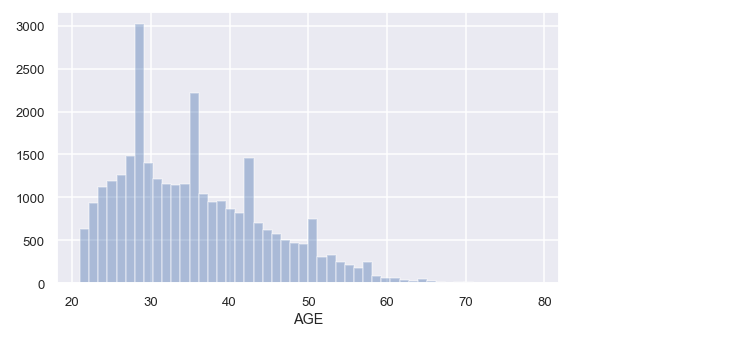


* Ta thấy có một bước nhảy lớn từ tháng 5,2005 của PAY\_5 đến tháng 7,2005 của PAY\_3 khi thanh toán chậm tăng đáng kể, sau đó đạt đỉnh ở tháng 8,2005 của PAY\_2, vài khoản bắt đầu tốt hơn ở tháng 9,2005 của PAY\_1.



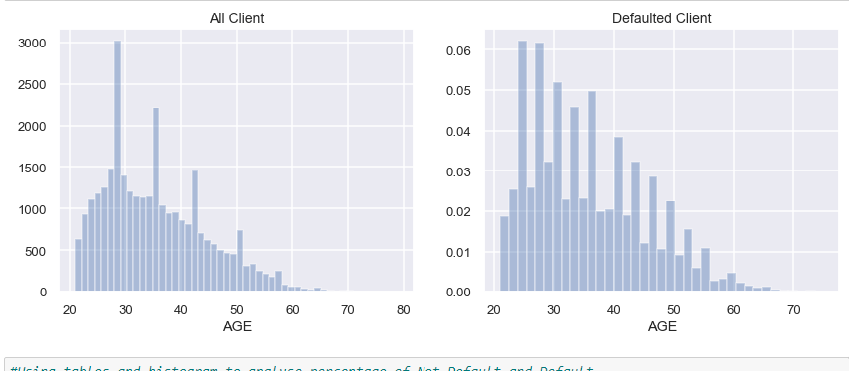
* Khi thanh toán bị trì hoãn hơn 2 tháng, các cơ hội của mặc định tăng 50%.

1. **Thống kê của cột AGE:**

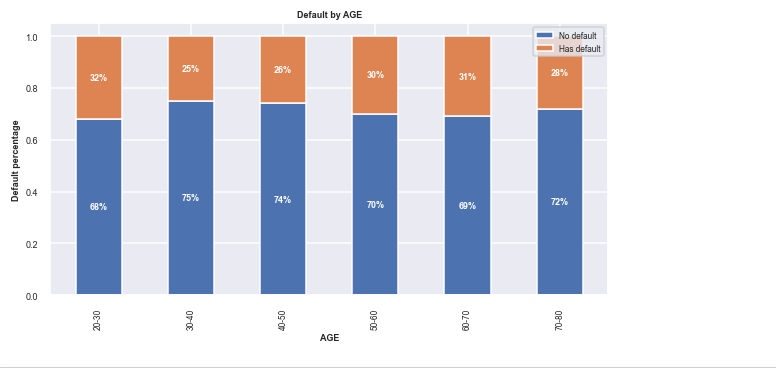


*Theo thống kê thấy thấy rằng:*

* Khách hàng trẻ nhất là 21 tuổi và lớn nhất là 79 tuổi. Hầu hết các khách hàng nằm ở độ tuổi từ 26 đến 35. Và một số nhóm tuổi cụ thể có số lượng người cao hơn.
* Biểu đồ so sánh 2 thành phần khách hàng: tất cả khách hang và khách hàng mặc định



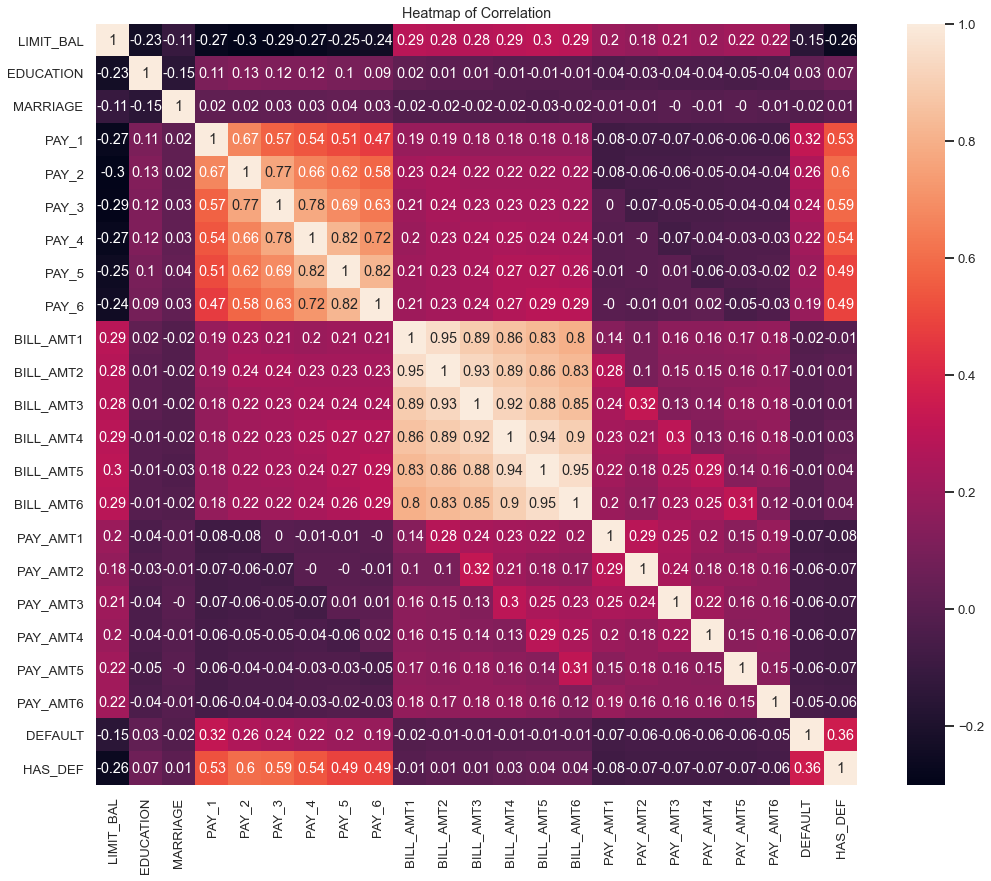
* Ta chia độ tuổi thành 6 nhóm: ‘20-30','30-40','40-50','50-60','60-70','70-80'.
* Với nhóm trên 60 ta có ít dữ liệu. Vì thế phần trăm có thể không tin cậy. Lấy một cái nhìn cụ thể ở khoảng tuổi từ 26 đến 36 thì ta có tỷ lệ Default thấp hơn so sánh.
* Xác suất default tăng cho khách hàng dưới 30 và trên 70 tuổi. Khách hàng ở độ tuổi 30-50 có tỷ lệ thanh toán trễ tháp nhất trong khi đó nhóm tuổi 20-30 và 50-70 có tỷ lệ thanh toán trễ cao hơn. Độ tuổi 30-50 là nguồn lực kiếm tiền chủ lực nhất trong xã hội.
* Tuy nhiên tỷ lệ trì hoãn thanh toán lại giảm nhẹ ở khách hàng độ tuổi trên 70 tuổi.



1. **Biểu đồ heatmap:**

*Từ biểu đồ heatmap có thể thấy*

* Các cột 'BILL\_AMT1', 'BILL\_AMT2', 'BILL\_AMT3', 'BILL\_AMT4', 'BILL\_AMT5', 'BILL\_AMT6' có mối tương quan cao với nhau, hay có thể nói nếu người đó sử dụng tiền thường không quá chênh lệch trong các tháng.
* Đối với các cột 'PAY\_1', 'PAY\_2', 'PAY\_3', 'PAY\_4', 'PAY\_5', 'PAY\_6' có mối tương quan tương đối cao với cột liền kề, hay nói khác hơn là tình trạng trả nợ cuar 2 tháng gần nhau có mối tương quan cao.
* Bên cạnh đó 'PAY\_1', 'PAY\_2', 'PAY\_3', 'PAY\_4', 'PAY\_5', 'PAY\_6' có mối tương quan cao với cột 'DEFAULT', nói cách khác thì tình trạng trả nợ có liên quan cao với các đối tượng lỗi.

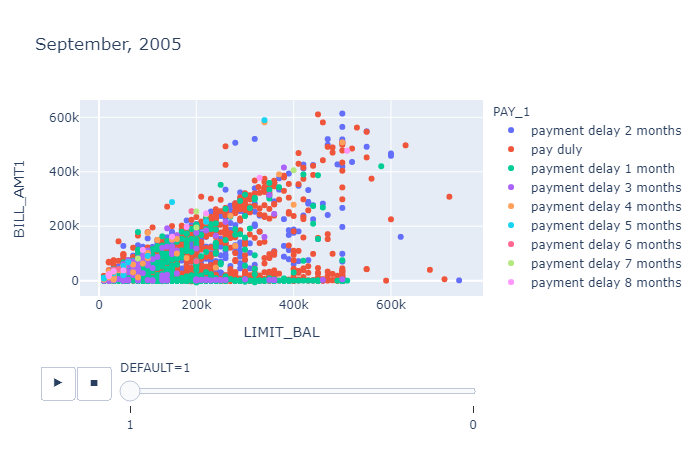


|  | **LIMIT\_BAL** | **EDUCATION** | **MARRIAGE** | **AGE** | **PAY\_1** | **BILL\_AMT1** | **PAY\_AMT1** | **DEFAULT** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **0** | 20000 | 2 | 1 | 20-30 | payment delay 2 months | 3913 | 0 | 1 |
| **1** | 120000 | 2 | 2 | 20-30 | pay duly | 2682 | 0 | 1 |
| **2** | 90000 | 2 | 2 | 30-40 | pay duly | 29239 | 1518 | 0 |
| **3** | 50000 | 2 | 1 | 30-40 | pay duly | 46990 | 2000 | 0 |
| **4** | 50000 | 2 | 1 | 50-60 | pay duly | 8617 | 2000 | 0 |

1. **Interact Visualization:**
2. **Figure 1:**

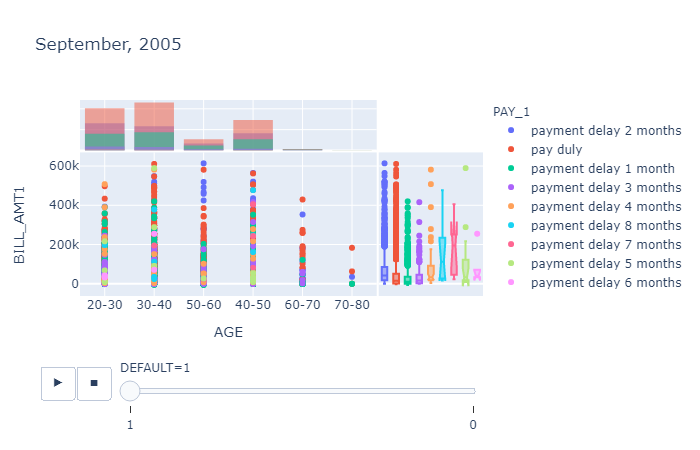
* Có nhiều giao dịch thanh toán hợp lệ.
* Đối với thanh toán trễ hạn thì đa số là trễ hạn từ 1-3 tháng.
* Dường như các giao dịch càng trễ hạn thì số tiền trả càng ít.
* Những đối tượng trong độ tuỏi từ 24-54 thường trả đúng hạn, và cũng sử dụng tiền tương đối nhiều.
* Càng về lớn tuổi có xu hướng ít sử dụng thẻ tín dụng hơn.

1. **Figure 2:**



* Càng lớn tuổi thì không có xu hướng sử dụng thẻ tín dụng.
* Người sử dụng thẻ có xu hướng sử dụng không vượt quá số tiền cho phép sử dụng.
* Đa số khách hàng sử dụng vượt quá số tiền cho phép thường là những khách hàng trả đúng hạn.

1. **Figure 3:**



* Đối tượng trên 70 tham gia sử dụng thẻ tín dụng khá ít và đa phần trả đúng hẹn nên có xu hướng không thanh toán vào tháng 10.
* Có vẻ khách hàng không thanh toán vào tháng tiếp theo nhiều hơn các khách hàng cần trả nợ vào tháng 10.
* Các khách hàng không thanh toán vào tháng tiếp theo thường trả đúng hạn, và không quá hạn quá lâu(2 tháng). Ngược lại các khác hàng lỗi có xu hướng kéo dài thời gian trả tiền.
* Các khách hàng cần trả nợ vào tháng tiếp theo có vẻ thường là đối tượng dưới 50 tuổi.

1. **Tóm lại:**

* Những Khách hàng không hoạt động không có nghĩa họ không có rủi ro mặc định. Ta tìm thấy 317 trong số 810 khách hàng không hoạt động không tiêu thụ trong 6 tháng sau đó đã mặc định trong tháng tới
* Để giảm rủi ro mặc định, phải rất thận trọng về hành vi thanh toán của khách hàng. Thận trọng hơn về khách hàng đã ly hôn và khách hàng cấp 3
* Một số dữ liệu chưa được giải thích cần lưu ý
* Một số khách hàng đã thanh toán đầy đủ trong tháng gần đây, mặc định.
* Một số khách hàng có số tiền hóa đơn âm. Nhưng vẫn sẽ được mặc định.
* Một số số tiền hóa đơn của khách hàng cao hơn nhiều so với hạn mức tín dụng của họ được cung cấp.
* Ta có thể đề xuất để quản lý một vài điều:
* Để giảm rủi ro mặc định, phải rất thận trọng về hành vi thanh toán của khách hàng.
* Thận trọng hơn về tình trạng khách hàng ly hôn và khách hàng cấp trung học.
* Chiến dịch tiếp thị nên nhắm đến độ tuổi của khách hàng từ 26 đến 45.
* Giao tiếp với nhóm thu thập dữ liệu.
* Đảm bảo tài liệu mô tả dữ liệu phù hợp.
* Giảm lỗi dữ liệu, kiểm tra hệ thống thu thập và theo dõi dữ liệu.