TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

KHOA TOÁN KINH TẾ

----------**

BÀI TẬP LỚN

MÔN: DATA-DRIVEN MARKETING

Đề tài: Phân tích chiến dịch quảng cáo mời khách hàng mở tài khoản tiết kiệm không kỳ hạn

Giảng viên: Nguyễn Quỳnh Giang

Lớp: DSEB 61

Nhóm: 2

Thành viên: 1. Bùi Thị Thu Hương

2. Nguyễn Thị Thảo Nguyên

3. Đỗ Thị Thanh Huế

4. Lê Tiến Bằng

Hà Nội, 2022

Phần 1: Description Analysis

1. Giới thiệu chung

Đây là tập dữ liệu gồm có 41188 dòng và 21 cột mô tả về kết quả các chiến dịch tiếp thị của ngân hàng. Chiến dịch thực hiện chủ yếu dựa trên các cuộc điện thoại trực tiếp, đề nghị khách hàng đăng kí gửi tiền có kỳ hạn. Nếu khách hàng đồng ý thì biến mục tiêu được đánh dấu là "Yes", ngược lại thì "No".

2. Dataset

- **Dữ liệu khách hàng của ngân hàng (Bank client data)**

1. age: Tuổi của khách hàng
2. job: Nghề nghiệp của khách hàng ('housemaid', 'services', 'admin.', 'blue-collar', 'technician', 'retired', 'management', 'unemployed', 'self-employed', 'unknown', 'entrepreneur', 'student')
3. marital: Tình trạng hôn nhân ('married', 'single', 'divorced', 'unknown')
4. education: Trình độ học vấn ('basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'professional.course', 'university.degree', 'unknown', 'illiterate')
5. default: Tình trạng thiếu nợ ('no','yes' ,'unknown')
6. housing: Có cho vay nhà ở không? ('no','yes','unknown')
7. loan: Có khoản vay cá nhân không? ('no', 'yes', 'unknown')

- **Liên quan đến người liên hệ cuối cùng của chiến dịch hiện tại**

8. contact: Phương thức liên hệ khách hàng ('cellular': điện thoại di động, 'telephone': điện thoại bàn)
9. month: Tháng gần nhất liên hệ với khách hàng ('jan', 'feb', 'mar', ..., 'nov', 'dec')
10. day_of_week: Ngày gần nhất liên hệ với khách hàng ('mon','tue','wed','thu','fri')
11. duration: Thời lượng liên lạc, tính bằng giây (số)
12. campaign: Số lần liên hệ được thực hiện trong chiến dịch
13. pdays: Số ngày trôi qua sau khi khách hàng được liên hệ lần cuối từ chiến dịch trước đó (số; 999 có nghĩa là khách hàng chưa được liên hệ trước đó)
14. previous: Số lần liên hệ được thực hiện trước chiến dịch này
15. poutcome: Kết quả của chiến dịch tiếp thị trước đó ('nonexistent', 'failure', 'success')

- **Thuộc tính bối cảnh kinh tế và xã hội (Economic Index attributes)**

16. emp.var.rate: Tỷ lệ thay đổi việc làm - chỉ số hàng quý
17. cons.price.idx: consumer price index - chỉ số giá tiêu dùng - chỉ số hàng tháng
18. cons.conf.idx: consumer confidence index chỉ số niềm tin của người tiêu dùng - chỉ số hàng tháng (số)
19. euribor3m: lãi suất 3 tháng của euribor - chỉ báo hàng ngày (số)

20. nr.employed: Số lượng nhân viên - chỉ số hàng quý

- **Biến đầu ra (mục tiêu mong muốn) - output**

21. y: Tình trạng đăng ký tiền gửi có kỳ hạn của khách hàng (no, yes)

3. Xử lý dữ liệu

```
[ ]  bank.isnull().sum().sum()
```

```
0
```

```
bank = bank.drop_duplicates()
bank.shape
```

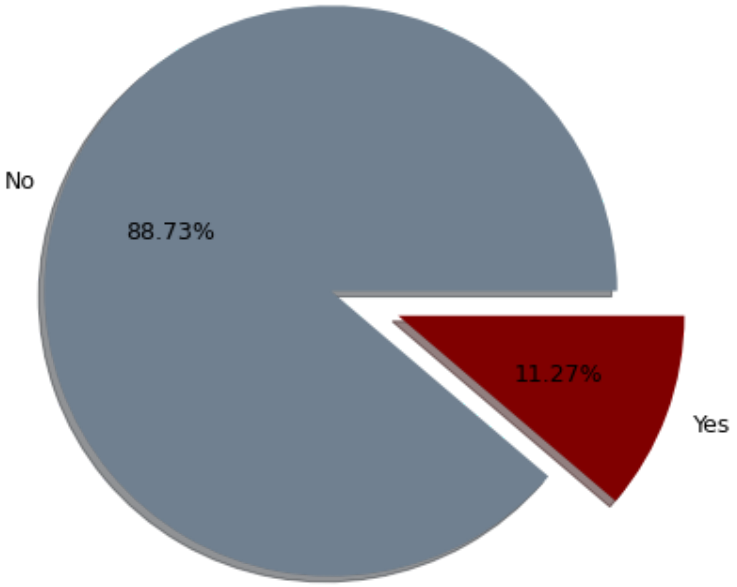
```
(41176, 21)
```

Dữ liệu không bị trống và có 12 dòng bị lặp. => Xóa bỏ các dòng bị lặp.

4. Tổng quan về chiến dịch

- **Conversion rate**

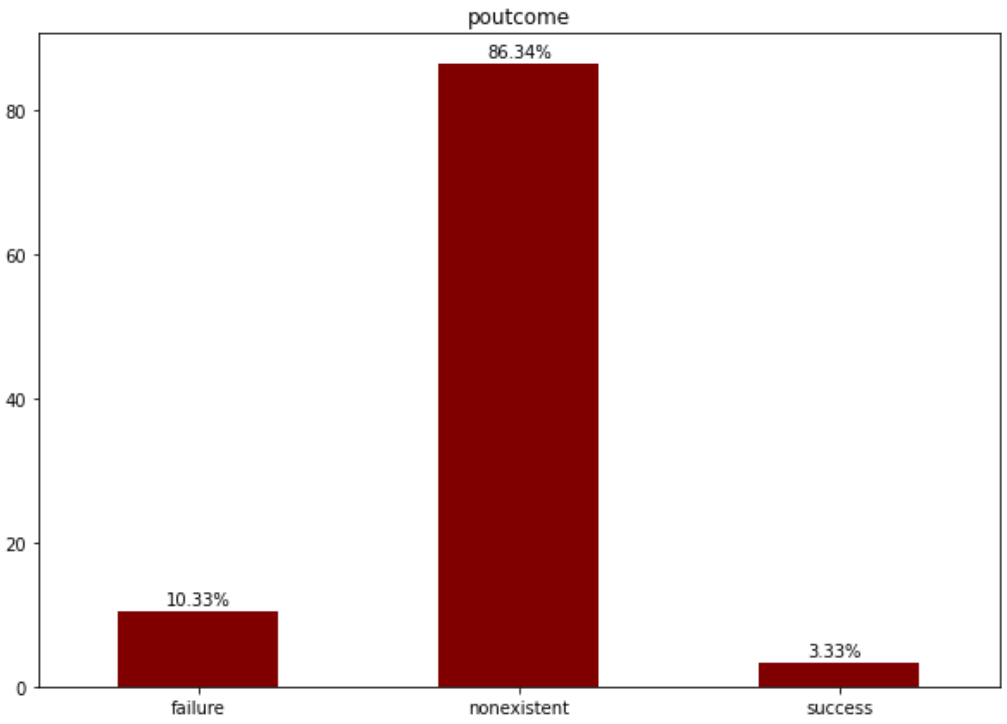
Theo kết quả của tập dữ liệu, công ty đã tiếp cận được 41176 khách hàng có độ tuổi từ 17 – 98 tuổi với nhiều ngành nghề khác nhau, trong đó có 4639 khách hàng đồng ý mở tài khoản có kỳ hạn, đạt tỉ lệ chuyển đổi là 11.27%.

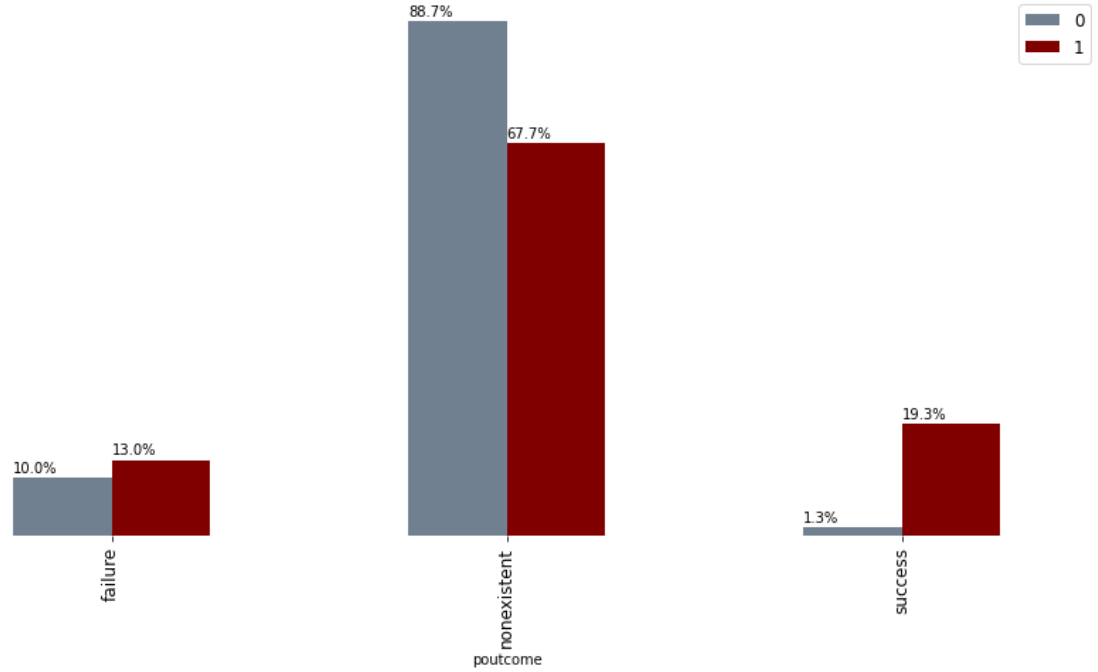


Tỷ lệ chuyển đổi khá thấp cho thấy chiến dịch chưa đạt hiệu quả cao.

- **Poutcome**

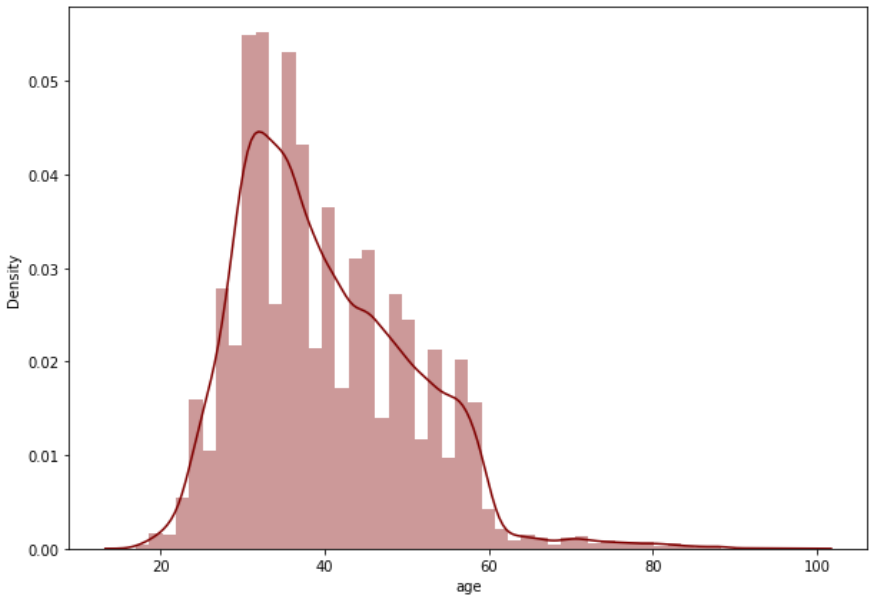
Nhìn vào biểu đồ bên dưới, chúng ta có thể thấy rằng phần lớn khách hàng được tiếp thị trong chiến dịch này đều là khách hàng chưa từng tham gia vào chiến dịch tiếp thị nào trước đó. Số lượng khách hàng mới này là 35551, chiếm 86.34% trên tổng số khách hàng tham gia. Có 10.33% khách hàng tham gia là khách hàng có kết quả thất bại và 3.33% khách hàng chấp nhận đăng kí mở tài khoản tiền gửi từ chiến dịch tiếp thị trước đó.





Mặc dù, tỉ lệ phần trăm khách hàng có kết quả "success" từ chiến dịch trước rất ít nhưng tỉ lệ khách hàng đó đồng ý đăng kí tài khoản có kỳ hạn lại cao hơn số khách hàng từ chối 18%. Số khách hàng mới được tiếp cận tuy nhiều nhưng tỉ lệ từ chối cũng rất cao, chiếm 88.7% trên tổng số khách hàng từ chối. Và tỉ lệ khách hàng chấp nhận mở tài khoản tiền gửi ở các nhóm 'failure', 'nonexistent', 'success' lần lượt là 13%, 67.7%, 19.3%.

- **Age**

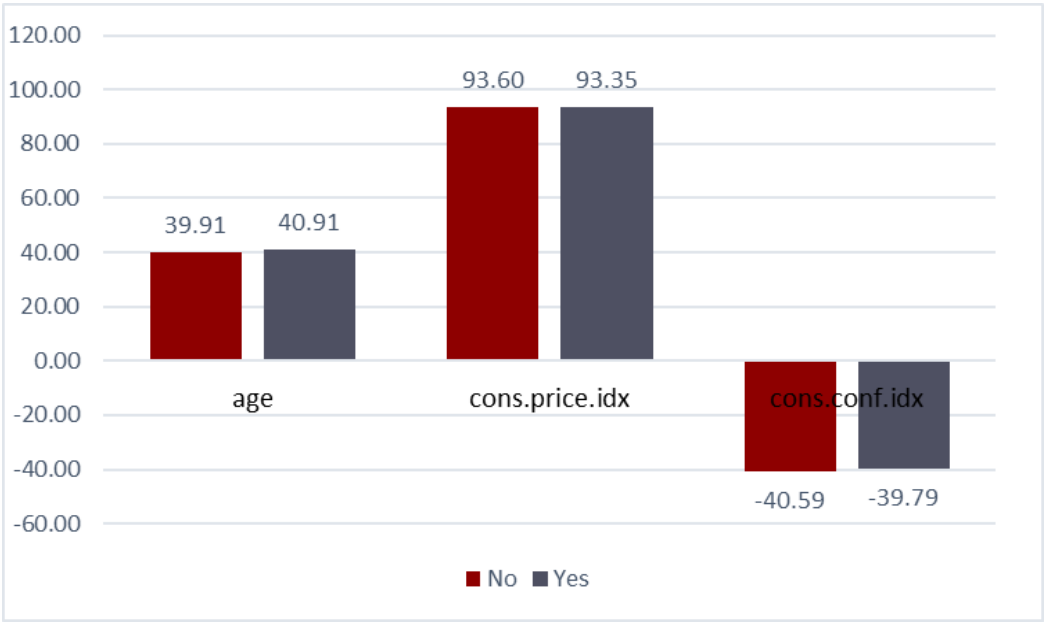


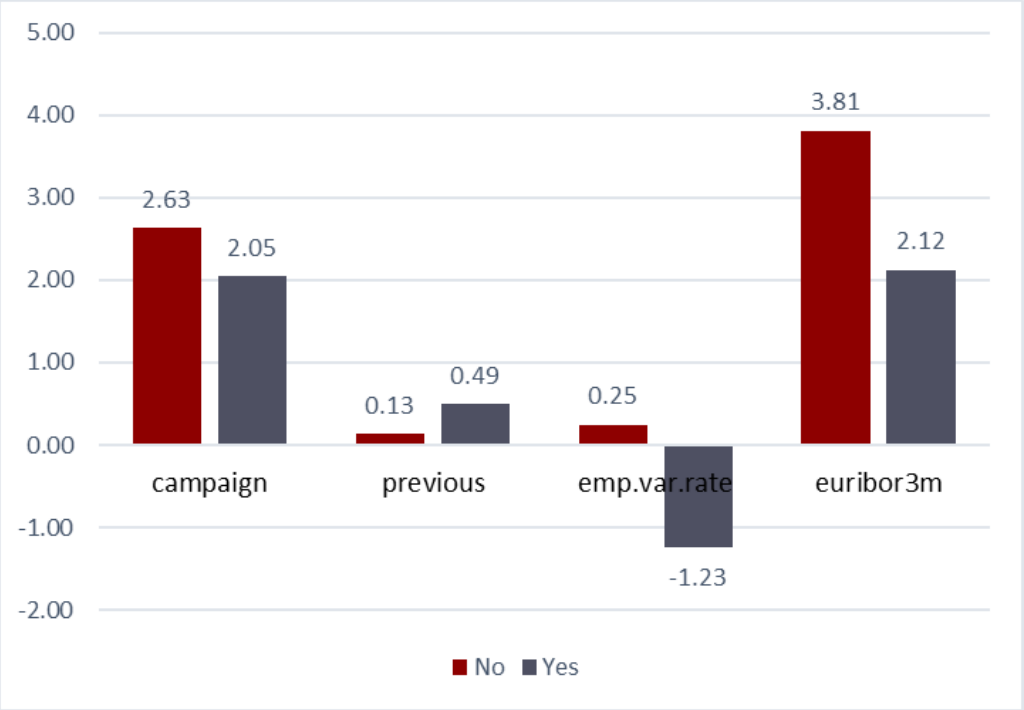
Nhìn vào đồ thị phân bổ độ tuổi, ta có thể rút ra nhận xét đa số khách hàng tham gia là người trưởng thành và đông nhất là ở độ tuổi khoảng 30+. Trong chiến dịch tiếp thị này, khách hàng có độ tuổi cao nhất là 98 và nhỏ nhất là 17 tuổi.

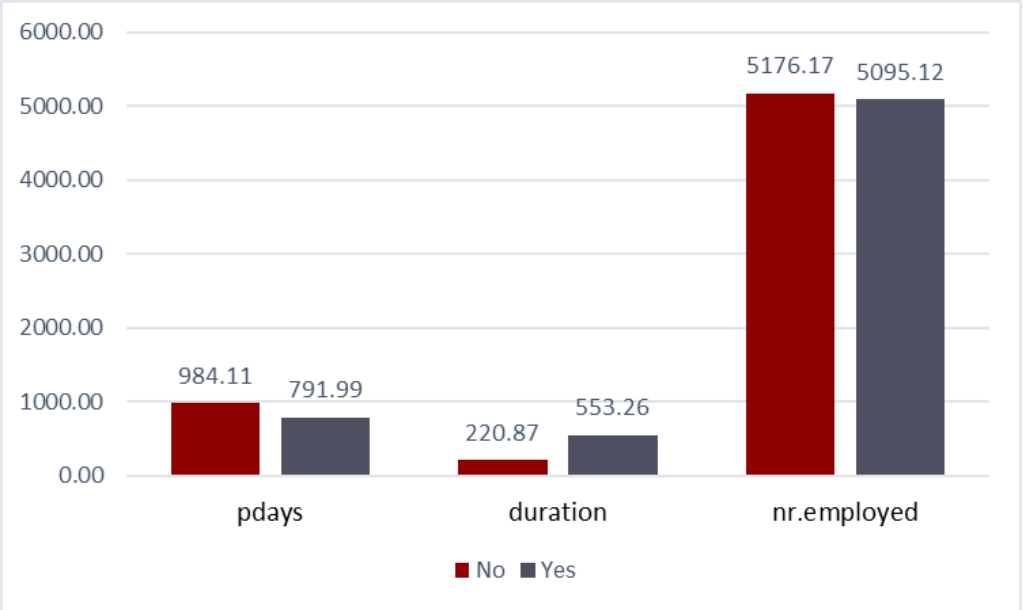
- **Mean**

```
bank.groupby('y').mean()
```

y	age	duration	campaign	pdays	previous	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
no	39.910994	220.868079	2.633385	984.109396	0.132414	0.248885	93.603798	-40.593232	3.811482	5176.165690
yes	40.912266	553.256090	2.051951	791.990946	0.492779	-1.233089	93.354577	-39.791119	2.123362	5095.120069







Từ bảng tính mean theo biến đầu ra 'y' và biểu đồ, ta có thể thấy được sự chênh lệch giữa các chỉ số và nhóm tôi rút ra được một số nhận xét như sau:

- Độ tuổi trung bình của khách hàng đăng ký tiền gửi có kỳ hạn cao hơn so với khách không đăng ký là 1 tuổi.
- Trung bình một khách hàng nhận được 2.6 cuộc gọi tiếp thị trong suốt chiến dịch. Tuy nhiên, thật ngạc nhiên khi khách hàng đã đăng ký tiền gửi có kỳ hạn nhận được số cuộc gọi thấp hơn.
- Trung bình tỷ lệ thay đổi việc làm ở nhóm đồng ý mở tài khoản là -1.23 trong khi nhóm không đăng ký có giá trị là 0.25
- Trung bình lãi suất gửi tín dụng ở khách hàng từ chối đăng ký là 3.81 và 2.12 ở nhóm đồng ý đăng ký.

- pdays: Số ngày kể từ khi khách hàng được liên hệ lần cuối của những khách hàng đã đăng ký tiền gửi thấp hơn so với khách hàng không đăng ký. Phải chăng, số ngày càng thấp, khả năng nhớ của khách hàng càng tốt và do đó cơ hội tiếp thị thành công cao hơn.
- Thời gian trung bình của các cuộc gọi tiếp thị là 258.3s. Dễ dàng nhận thấy thời gian trung bình của các cuộc gọi có khách hàng đồng ý đăng ký lớn hơn so với các cuộc gọi tiếp thị thất bại.

=> Thời lượng cuộc gọi càng cao, khách hàng càng có nhiều khả năng mở một khoản tiền gửi có kỳ hạn. Vì thời lượng cuộc gọi là đặc điểm tương quan tích cực nhất đến việc liệu khách hàng tiềm năng có mở một khoản tiền gửi có kỳ hạn hay không, bằng cách đưa ra những câu hỏi thú vị cho khách hàng trong suốt cuộc gọi nên thời lượng cuộc trò chuyện có thể tăng lên. Điều này không đảm bảo rằng khách hàng sẽ đăng ký tuy nhiên, chúng ta không nên bỏ sót đặc điểm này.

Action: Vì vậy, bằng cách thực hiện chiến lược tăng mức độ tương tác với khách hàng tiềm năng => tăng xác suất đồng ý đăng kí => ngân hàng sẽ tăng hiệu quả cho chiến dịch tiếp thị tiếp theo.

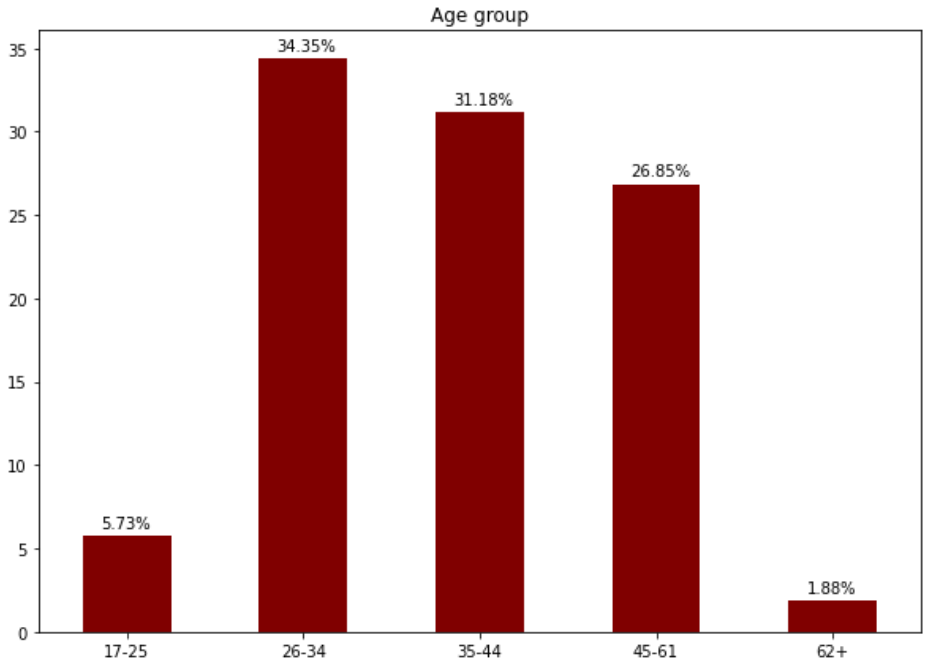
- Trung bình một quý ở ngân hàng có 5095 nhân viên tiếp thị thực hiện tiếp thị thành công đến khách hàng.

5. Insight:

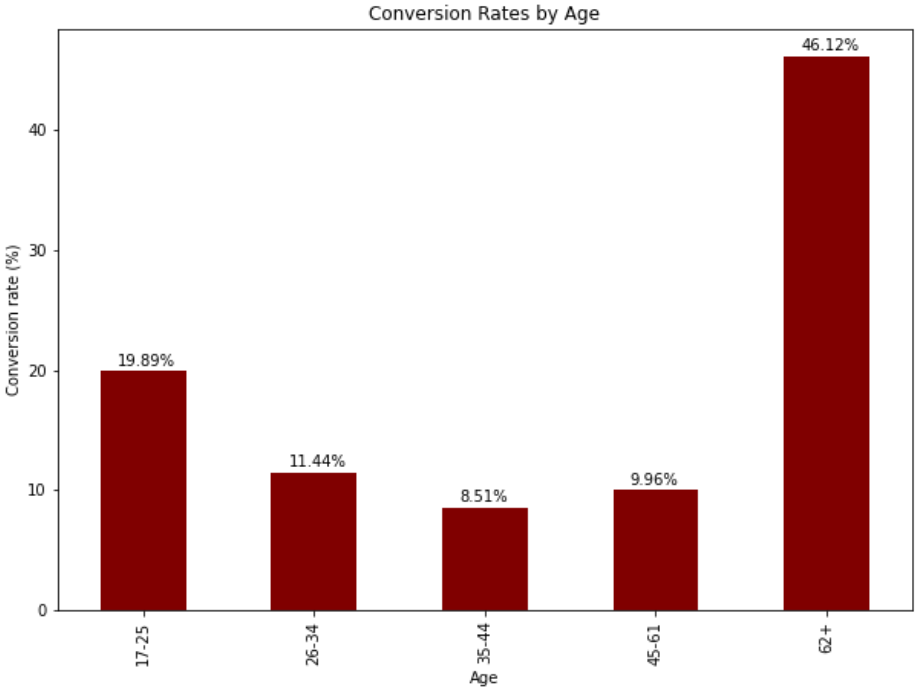
- **Conversion rate by Age**

Để dễ cho việc quan sát và đánh giá chất lượng của chiến dịch, chúng tôi chia dữ liệu theo 5 nhóm tuổi như sau:

- 17-25: Học sinh, sinh viên hoặc người mới đi làm
- 26-34: Nhóm lao động trẻ
- 35-44: Nhóm lao động có kinh nghiệm
- 45-61: Nhóm lao động giàu kinh nghiệm
- 62+ : Người làm việc quá tuổi hoặc người đã nghỉ hưu



Quan sát biểu đồ, ta thấy rằng nhóm lao động trẻ 17-25 và nhóm lao động có kinh nghiệm 35-44 là 2 nhóm khách hàng tham gia chiến dịch nhiều nhất với hơn 65%, nhóm người làm việc quá tuổi hoặc đã nghỉ hưu 62+ là nhóm có số lượng tham gia ít nhất với 1.88%.



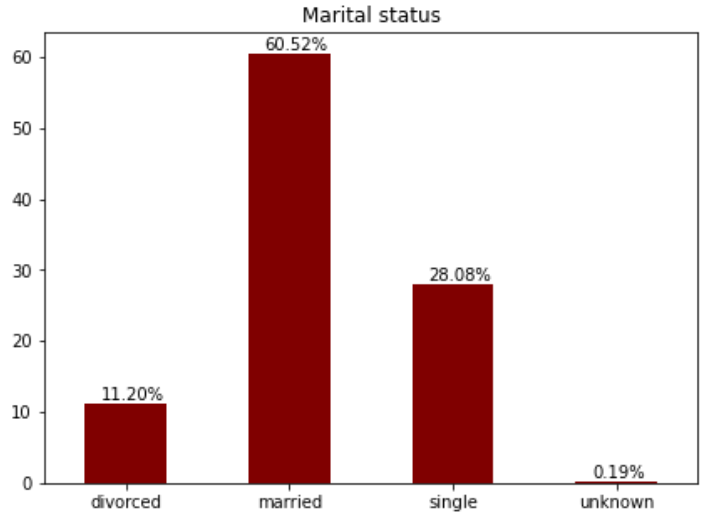
Có thể thấy rằng mọi người phần lớn không đồng ý mở tài khoản khi ngân hàng liên hệ với họ về vấn đề này bởi 3 nhóm có đông người tham gia chiến dịch nhất có tỉ lệ chuyển đổi rất thấp. Mặc dù, nhóm tuổi 62+ tham gia chiến dịch rất ít nhưng tỉ lệ chuyển đổi rất cao, đạt 46.12%. Có lẽ, do vấn đề tuổi cao nên họ muốn tìm một nơi an toàn giữ tiền để an dưỡng tuổi già. Nhóm khách hàng trong độ tuổi 17-25 đạt tỉ lệ

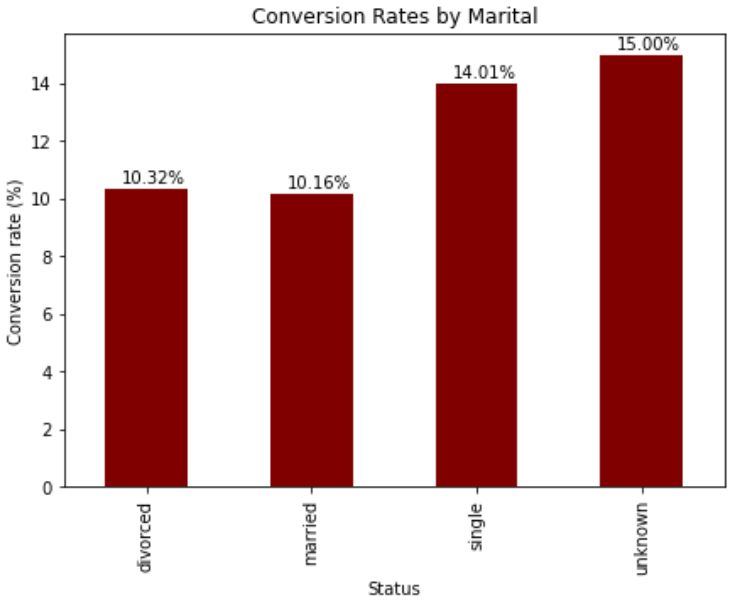
chuyển đổi ⅕, đây có thể là do họ là còn đi học hoặc mới tiến vào thị trường lao động nên chưa có nhiều tiền để đầu tư vào các danh mục sinh lời cao hơn, vì vậy họ đồng ý mở tài khoản.

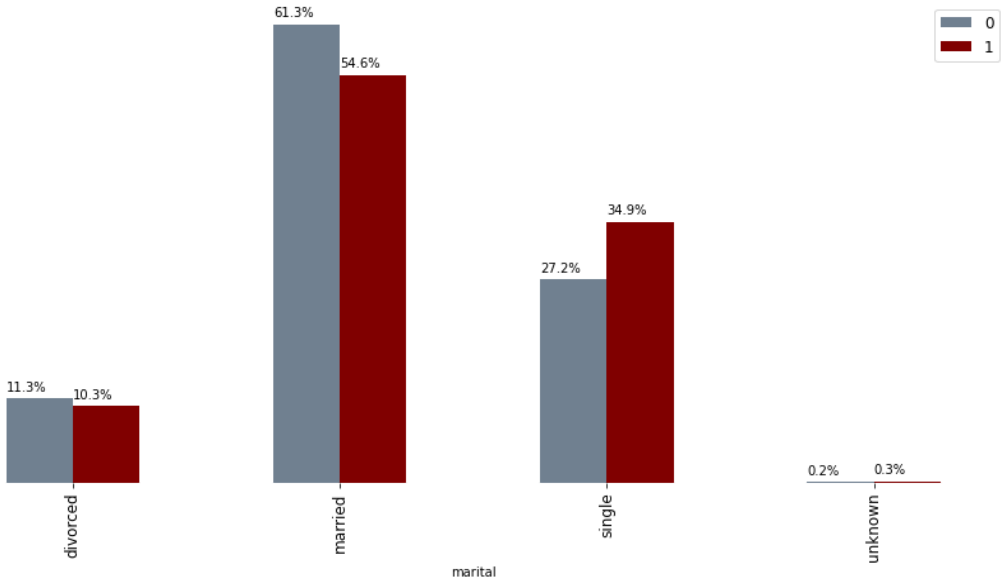
Action: Nên tập trung tiếp thị vào 2 nhóm tuổi 17-25 và 62+ bởi họ là nhóm có nhu cầu cao hơn các nhóm khác. Để thu hút thêm khách hàng, ngân hàng cần có chiến dịch marketing thú vị, chất lượng dịch vụ cũng như các chương trình tri ân, chế độ hậu mãi tốt hơn.

- **Conversion rate by Marital**

Vì khách hàng chủ yếu của chiến dịch tiếp thị này ở độ tuổi 26-60 tuổi nên hầu hết họ đều đã kết hôn, chiếm 60.52% tổng số khách hàng tham gia, tuy nhiên đây lại là nhóm có tỉ lệ chuyển đổi thấp nhất với 10.16%. Nhóm unknown tuy chỉ chiếm 0.19% lượng khách hàng nhưng tỉ lệ chuyển đổi cao nhất - 15%, đây có thể là do số lượng khách hàng nhóm này quá ít.



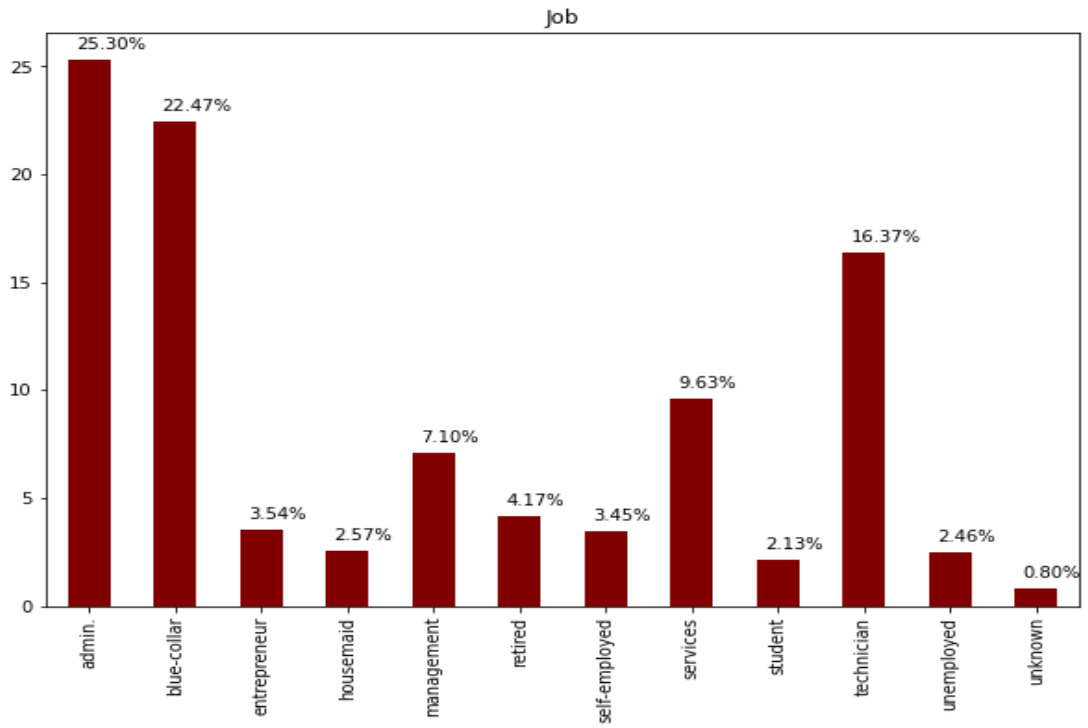




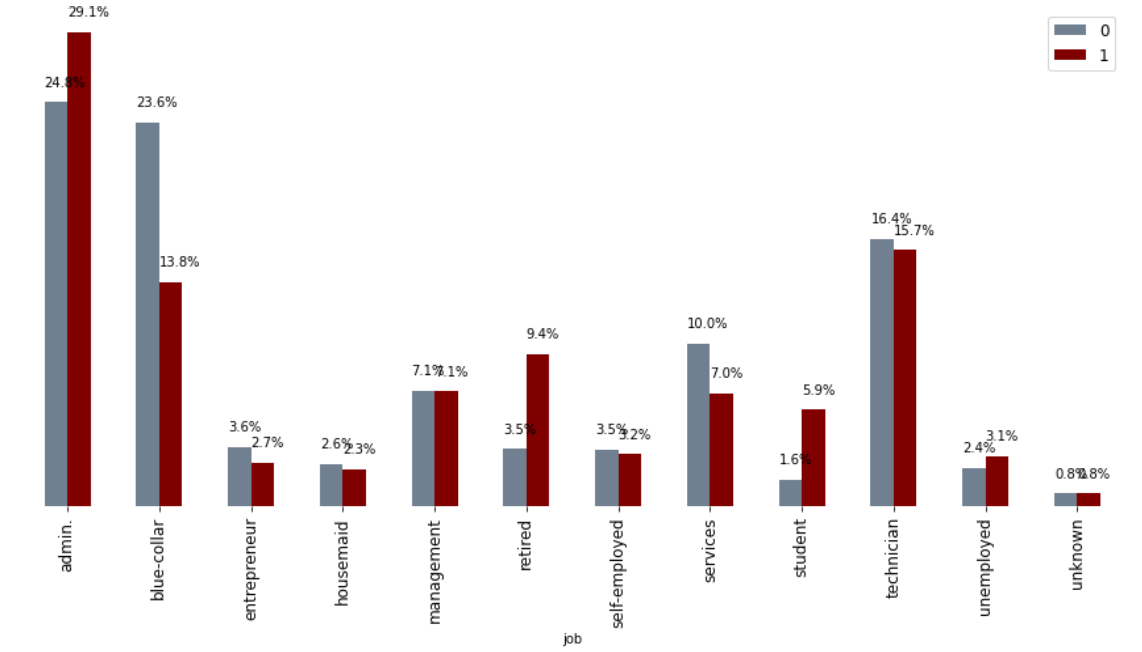
Ở đây, chúng ta có thể thấy nhóm đã kết hôn có nhiều người đăng ký nhất, chiếm 54,6% tổng số người đăng ký, nhiều hơn tổng số người từ nhóm độc thân và ly hôn (lần lượt là 34,9% và 10,3%). Tuy nhiên, nếu xét riêng từng nhóm thì nhóm độc thân có khả năng đăng kí cao hơn là từ chối, còn nhóm đã kết hôn có tỉ lệ từ chối cao hơn. Có lẽ là do họ phần lớn là nhóm đã tham gia vào thị trường lao động lâu năm nên có nhiều kiến thức cũng như tài sản đem đi đầu tư vào những danh mục sinh lời nhiều hơn hoặc do họ đã có con cái nên họ để tiền nuôi con, trang trải cuộc sống.

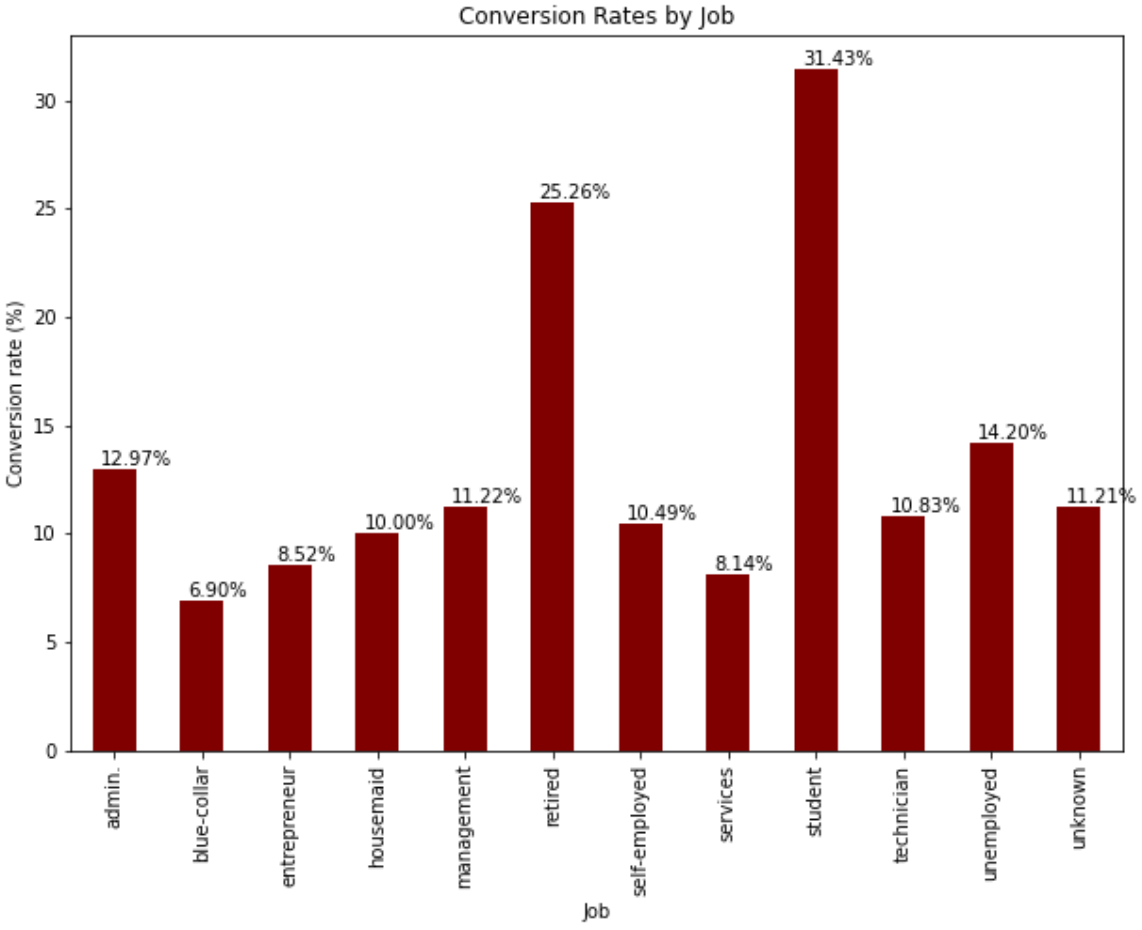
Action: Đưa ra nhiều ưu đãi phù hợp với nhu cầu của nhóm khách hàng đã kết hôn để tăng lượng khách mở tài khoản.

- **Do job titles affect conversion rate?**



Dễ dàng nhận thấy khách hàng tham gia chiến dịch của ngân hàng tập trung chủ yếu ở nhóm ngành nghề ‘admin’ (25.30%), ‘blue-collar’ (22.47%) và ‘technician’ (16.36%). Chỉ khách hàng đến từ 3 nhóm nghề này đã chiếm 64.13% lượng người tham gia. Và 2 nhóm có ít người tham gia nhất là ‘student’ và ‘unknown’ với giá trị lần lượt là 2.13% và 0.8%.



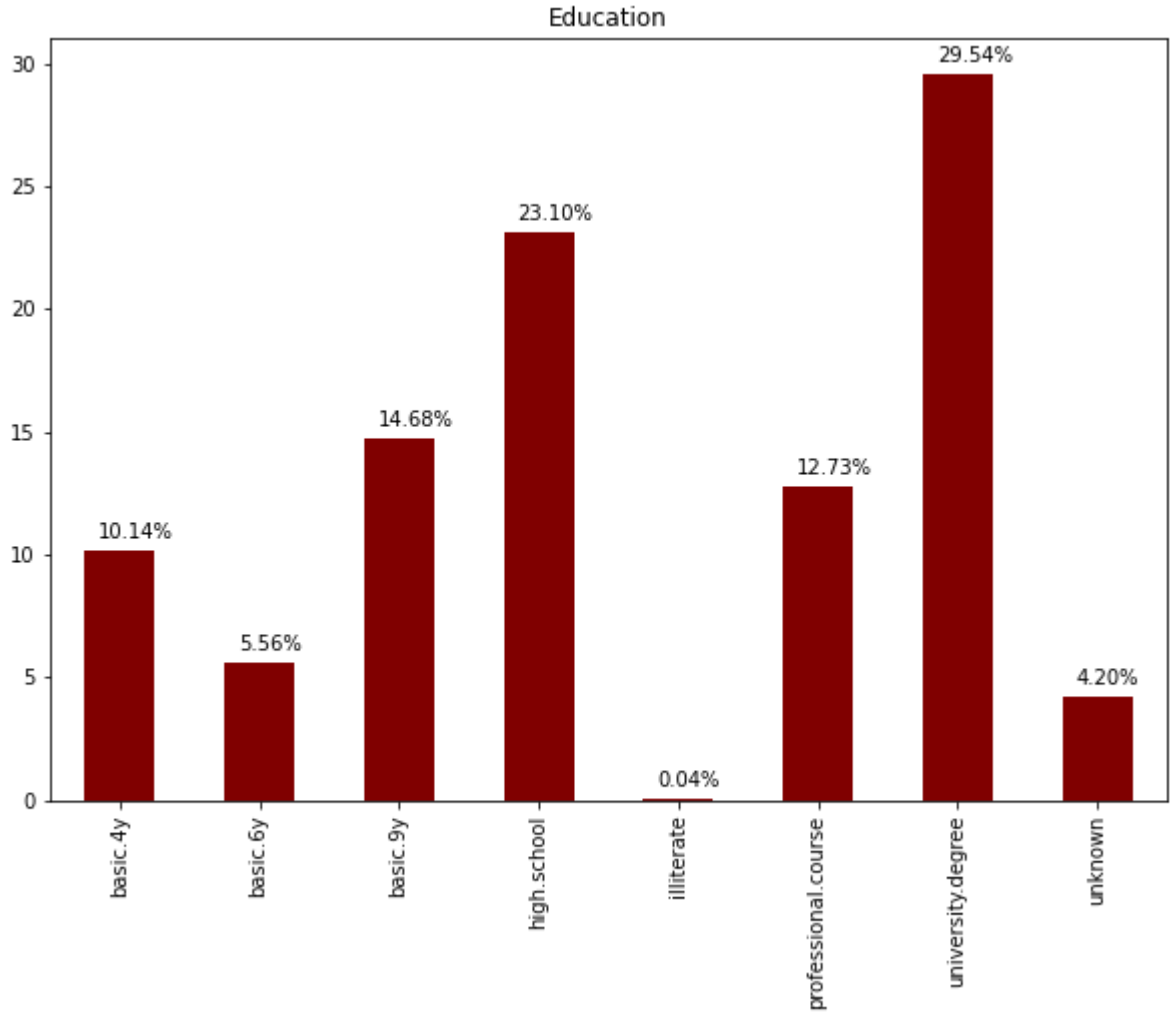


Tỉ lệ chuyển đổi ở nhóm khách hàng Học sinh và Người đã nghỉ hưu là cao nhất với giá trị lần lượt là 31.43% và 25.26% trong khi họ không phải nhóm khách hàng tham gia chiến dịch lớn nhất. Nhóm Học sinh-sinh viên có lẽ do còn đi học hoặc mới tiếp xúc với thị trường lao động nên chưa có nhiều tiền để đầu tư vào các danh mục sinh lời cao hơn nên họ chọn cách gửi tiền, tuy lãi suất thấp nhưng đó là khoản đầu tư an toàn nhất. Và điều đó cũng tương tư với nhóm Người đã nghỉ hưu, họ đã đến tuổi nghỉ ngơi, an hưởng tuổi già nên họ chọn mở tài khoản để cất trữ tiền cho cuộc sống, cho con cháu. Các nhóm ngành nghề khách có tỉ lệ chuyển đổi tương đối đồng đều. Nhóm khách hàng công nhân và người làm dịch vụ có tỉ lệ chuyển đổi thấp nhất. Có lẽ thu nhập của nhóm này không được cao nên nó ảnh hưởng đến quyết định đăng kí của họ.

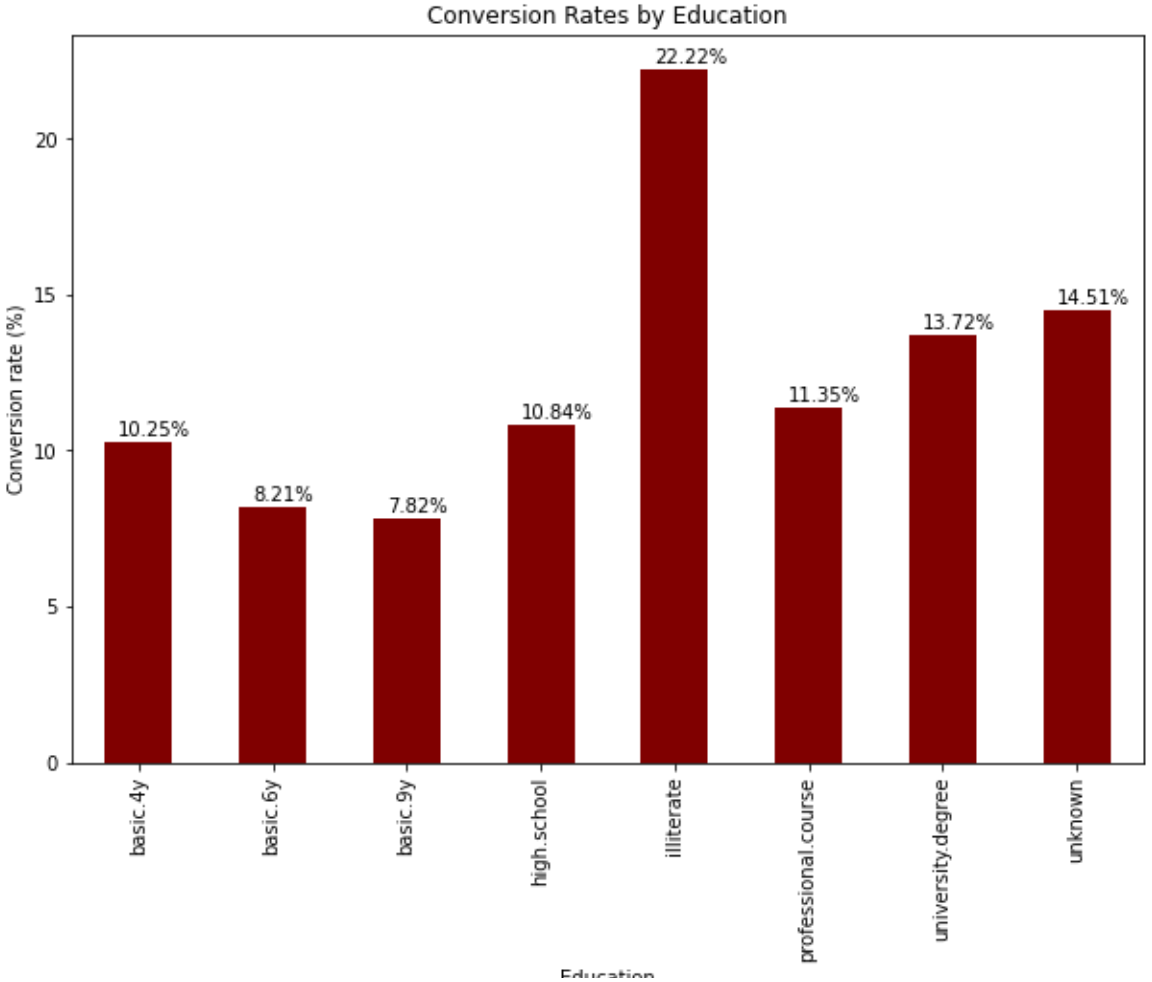
=> Job có ảnh hưởng tới conversion rate.

Action: Đẩy mạnh tiếp thị ở nhóm khách hàng đã nghỉ hưu và nhóm là học sinh, sinh viên. Chúng ta cũng có thể tiếp cận nhóm khách hàng này tại các câu lạc bộ người cao tuổi hoặc các trường học.

- **Does education affect conversion rate?**

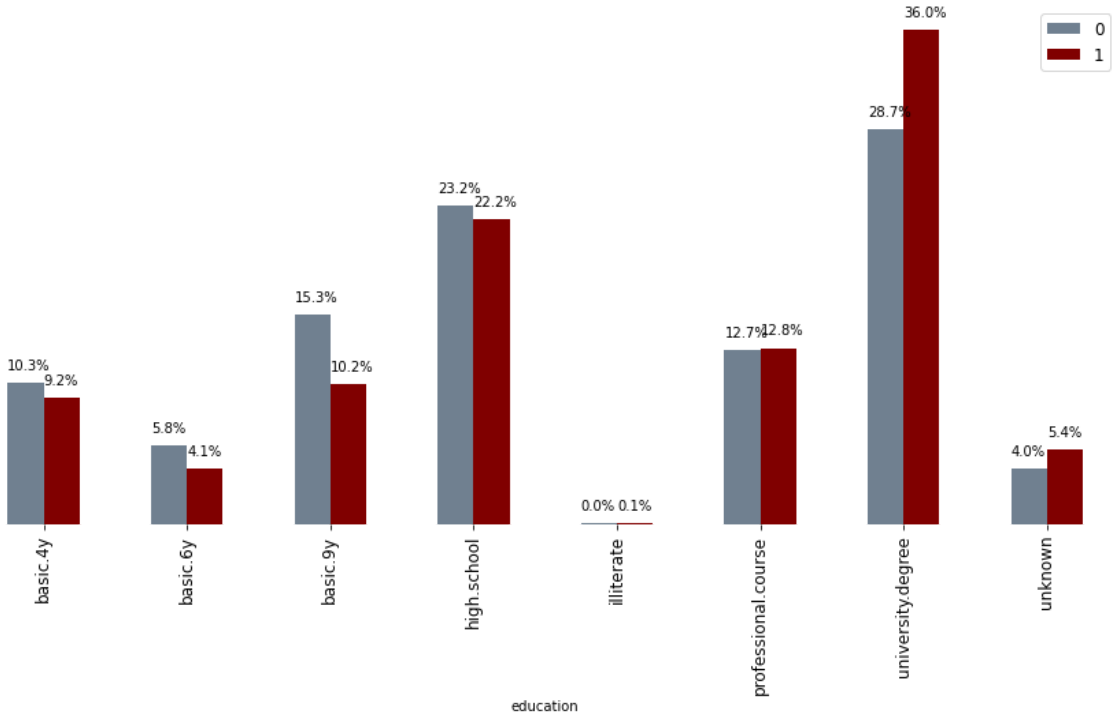


Nhóm có trình độ học vấn ‘university.degree’ và ‘high.school’ là tệp khách hàng được tiếp thị nhiều nhất trong chiến dịch lần này với hơn 50%. Nhóm ‘illiterate’ là nhóm có ít khách hàng nhất, số lượng người ‘mù chữ’ tham gia vào chiến dịch lần này là 18.



Biểu đồ cho thấy nhóm ‘illiterate’ có tỉ lệ chuyển đổi cao nhất, chiếm 22.22% tuy nhiên số lượng khách hàng thuộc nhóm này quá thấp nên không thể đánh giá chính xác được. Các nhóm trình độ còn lại có tỉ lệ chuyển đổi xấp xỉ nhau, không vượt quá 15% và chênh nhau không quá 7%.

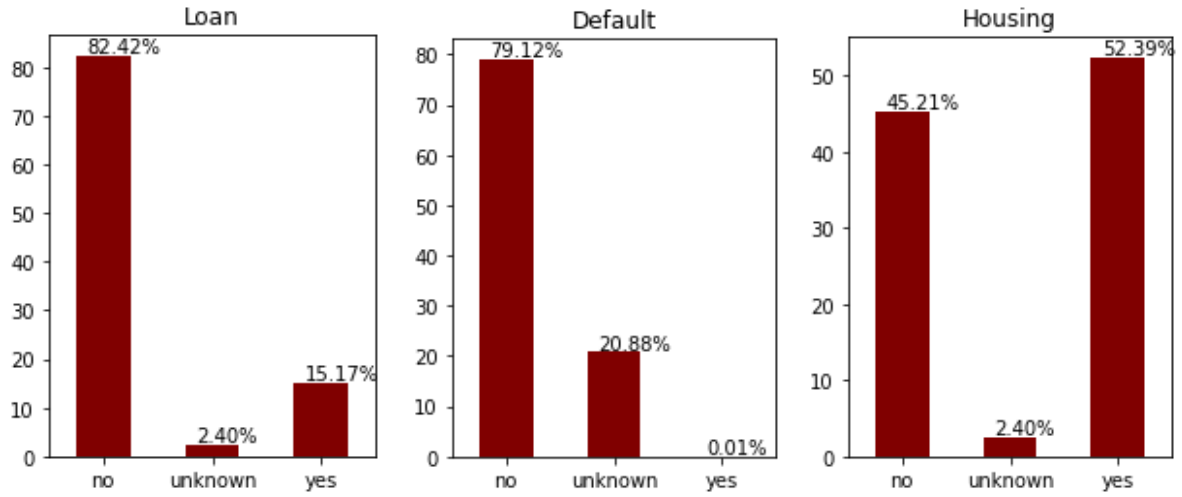
=> Trình độ học vấn của khách hàng không có ảnh hưởng nhiều đến conversion rate.



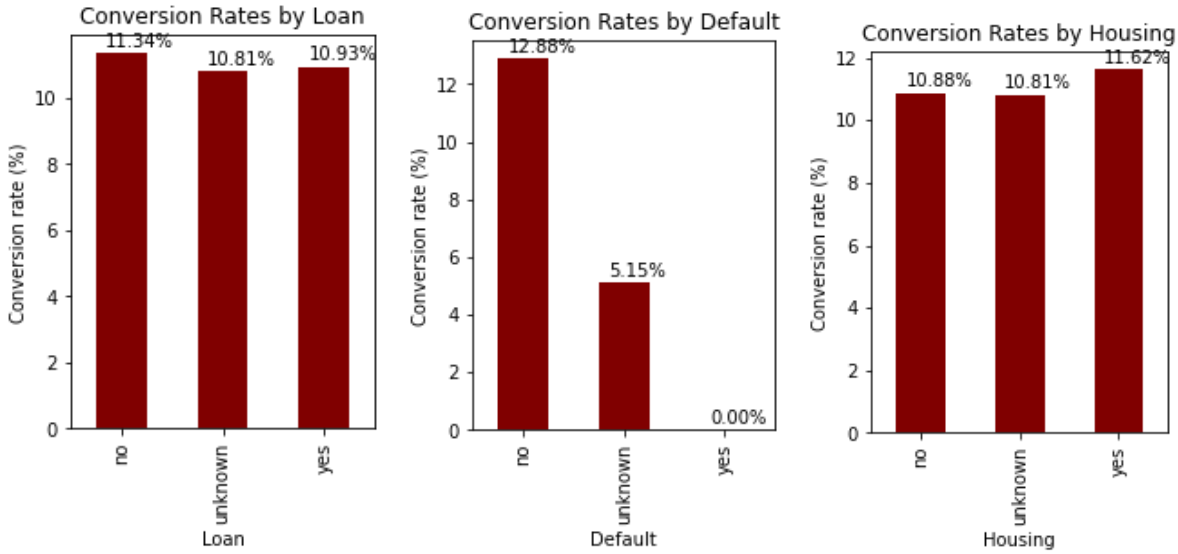
Có thể thấy rằng, nhóm khách hàng có trình độ ‘university.degree’ có xu hướng chấp nhận mở tài khoản cao hơn là từ chối.

Action: Thực hiện chiến dịch marketing ở nhóm khách hàng có trình độ học vấn này.

- **Is the conversion rate higher in the group having loan, default & housing?**



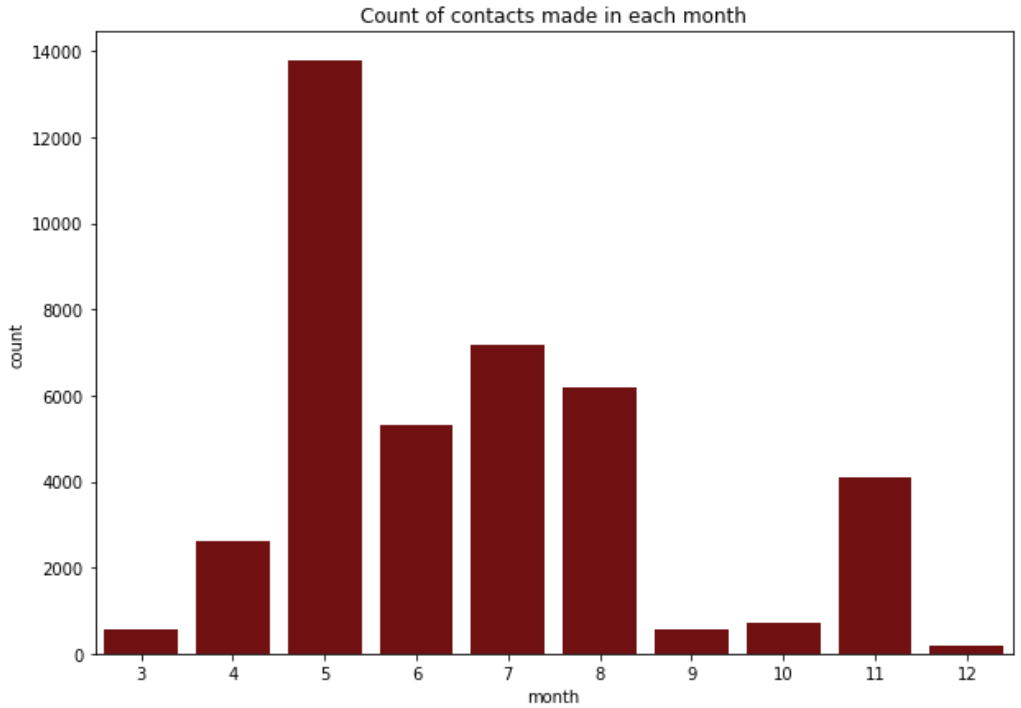
Khách hàng tham gia chiến dịch tiếp thị này của ngân hàng hầu như không có các khoản vay cá nhân hay khoản vay tín dụng, riêng khoản vay mua nhà thì số lượng giữa có và không khá tương đồng.



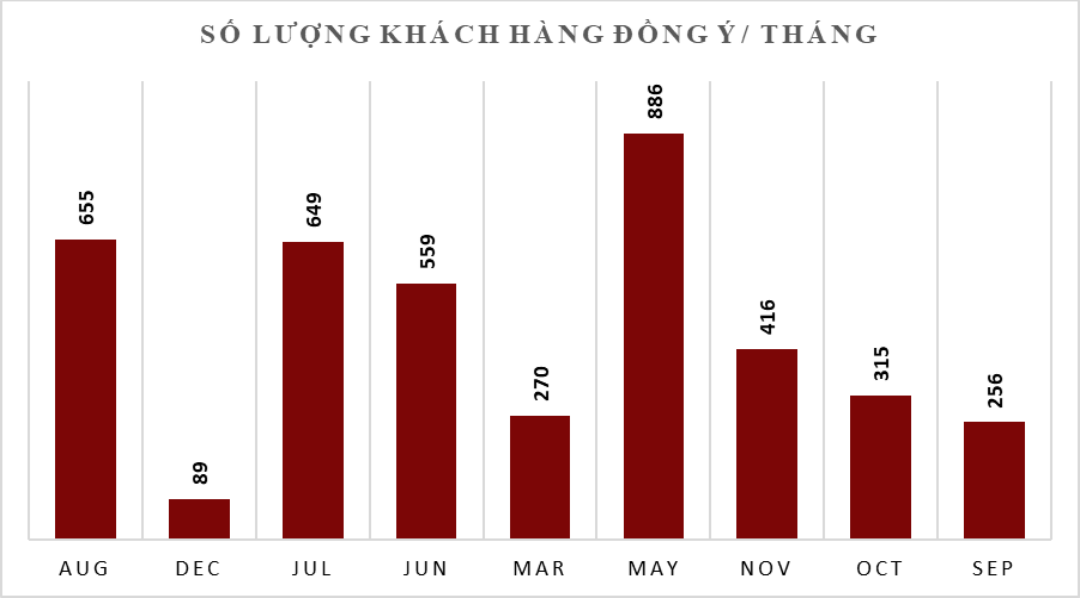
Tỉ lệ chuyển đổi theo khoản nợ cá nhân, khoản vay mua nhà giữa các nhóm khách hàng là tương đối cao đồng đều, có chút chênh lệch nhưng không đáng kể. Riêng tỉ lệ chuyển đổi theo khoản vay tín dụng thì có sự chênh lệch lớn, nhóm khách hàng có khoản vay tín dụng sẽ từ chối mở tài khoản (0%).

Action: Nếu khách hàng có khoản vay tín dụng thì loại bỏ khỏi danh sách tiếp thị, tránh lãng phí thời gian, nhân lực.

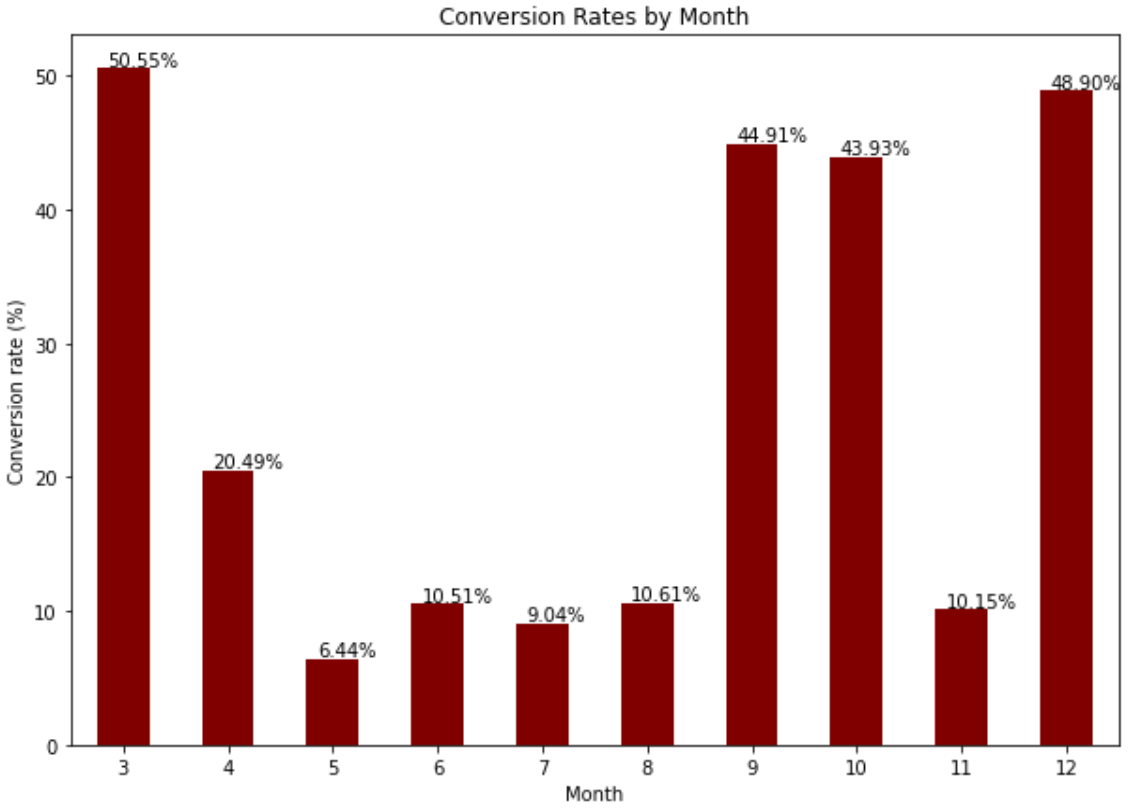
- **Which month should we contact our customer?**



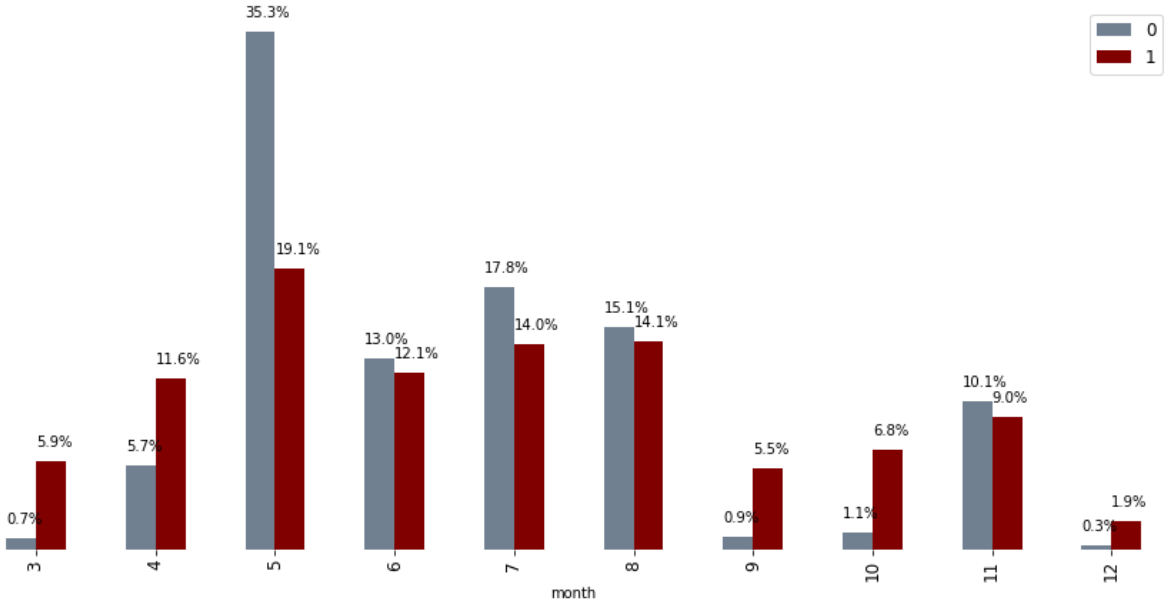
Nhìn chung, ta có thể thấy tháng 5 là tháng có số lượng cuộc gọi nhiều nhất (13767 cuộc gọi ~ 33.42%) và cao gần gấp đôi tháng 7 - tháng có số lượng cuộc gọi cao thứ 2. Trong đó tháng 12 là tháng ít liên lạc với khách hàng nhất, chiếm chưa tới 1% (cụ thể là 0.44%) số lượng cuộc gọi trong toàn chiến dịch.



Vì tháng 5 là tháng có số lượng cuộc gọi tới khách hàng nhiều nhất, nên số lượng khách hàng đồng ý/ tháng của tháng 5 cũng đạt nhiều nhất với 886 khách hàng đồng ý mở tài khoản tiền gửi có kì hạn. Tháng 7 và tháng 8 cũng đạt được số lượng khách hàng đồng ý khá ấn tượng với hơn 600 khách hàng 'gật đầu'. Tháng 12, ngân hàng có thêm 89 vị khách mới.



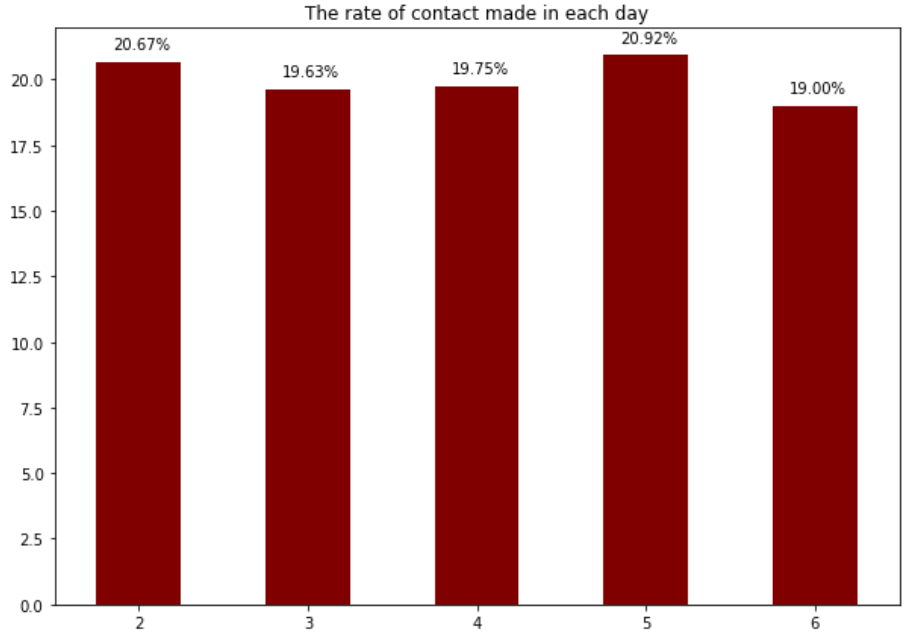
Tuy tháng 5 là tháng có số lượng khách hàng đồng ý mở tài khoản tiết kiệm cao nhất nhưng tỉ lệ chuyển đổi lại thấp nhất, chỉ có 6.44%. Tháng 12, tháng 3, tháng 9 và tháng 10 có tỉ lệ chuyển đổi rất cao với giá trị lần lượt là 50.55%, 48.9%, 44.91% và 43.93% trong khi số lượng liên hệ của các tháng này thấp nhất.



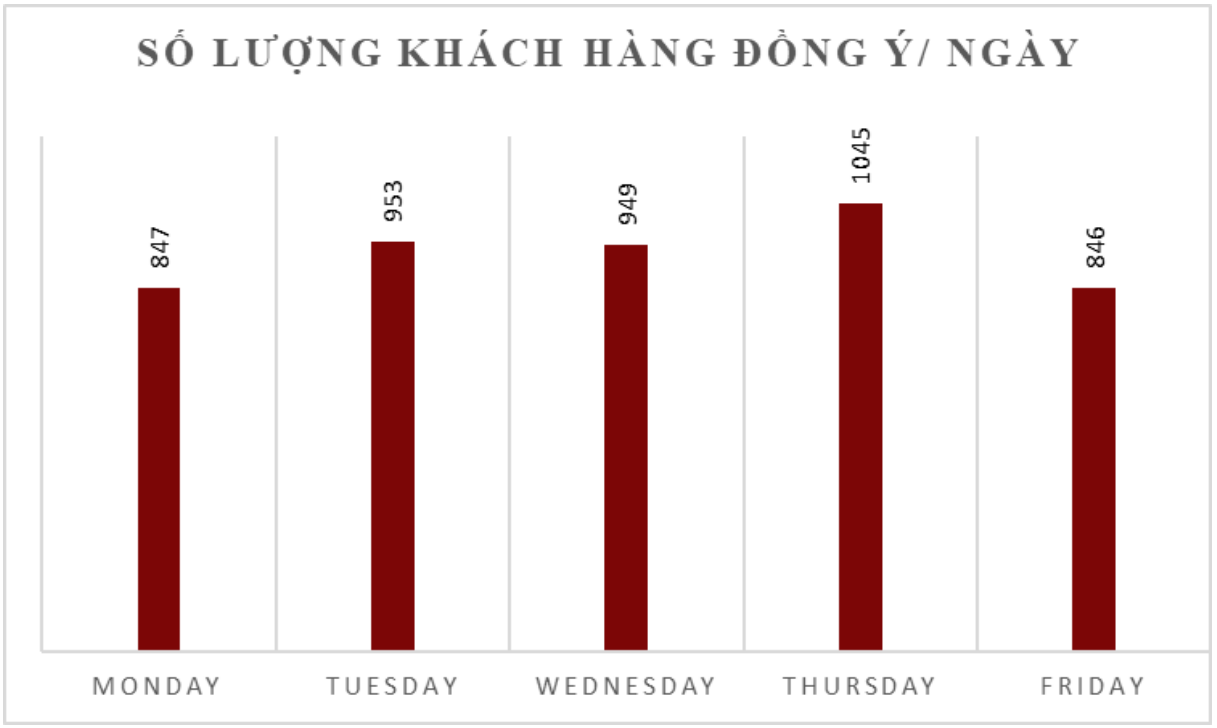
Tháng 4 có tỉ lệ chuyển đổi là 20.49% nhưng phần trăm khách hàng đăng kí tài khoản của tháng 4/ tổng số khách hàng đăng kí nhìn chung cao hơn % khách hàng không đồng ý.

Action: Các chiến dịch tiếp thị tiếp theo nên tập trung hoạt động, liên hệ với khách hàng vào các tháng 3, 9, 10 và 12 và ngân hàng cũng nên tìm hiểu nguyên nhân tại sao các tháng hoạt động trọng điểm lại không đạt được kết quả cao, từ đó đưa ra giải pháp khắc phục.

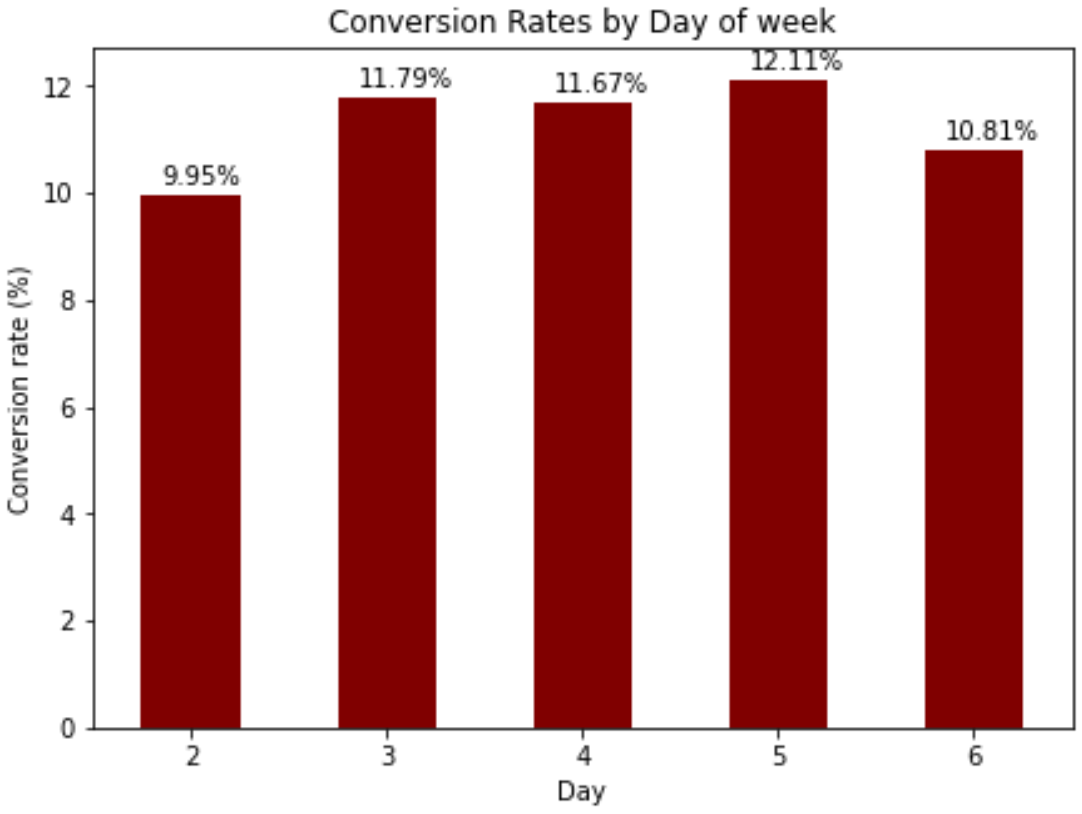
- **Which day should we contact our customer?**

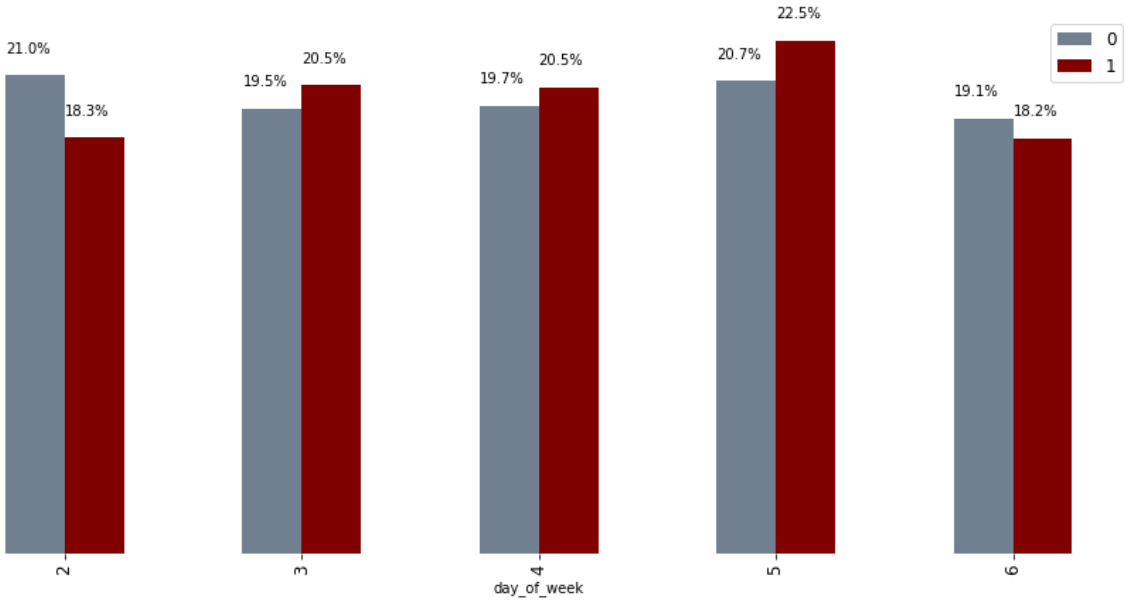


Nhìn vào biểu đồ, dễ dàng nhận thấy rằng tỉ lệ các cuộc gọi các ngày làm việc trong tuần khá đều nhau, có sự chênh lệch không đáng kể.



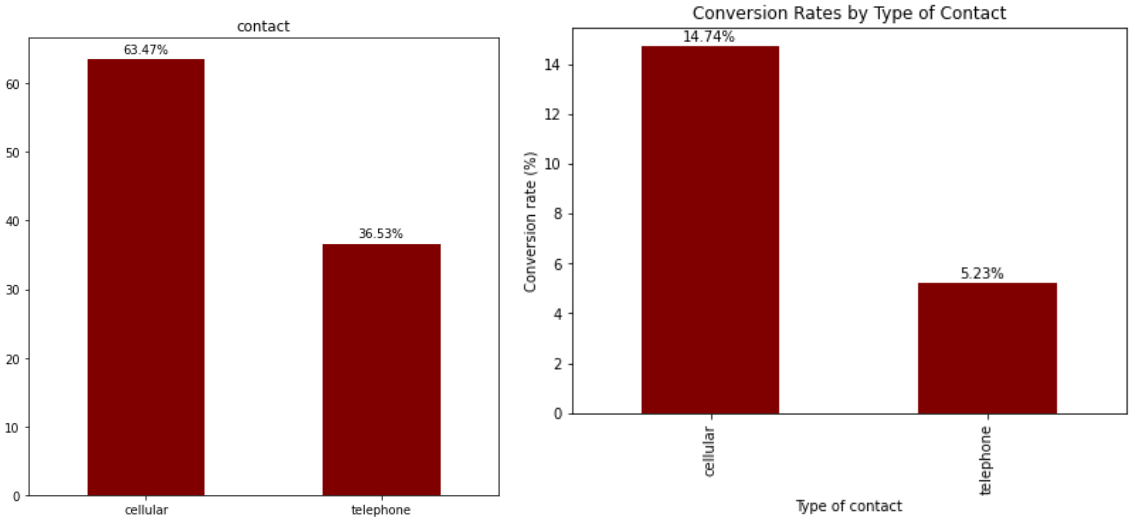
Theo như kết quả của tệp dữ liệu, thứ 5 là ngày có số lượng khách hàng chấp nhận mở tài khoản nhiều nhất - 1045 khách hàng. Tuy nhiên, các ngày còn lại cũng có số lượng khách hàng đồng ý rất cao, khá đồng đều và thứ 6 là ngày có ít khách hàng đồng ý nhất với 846 người.



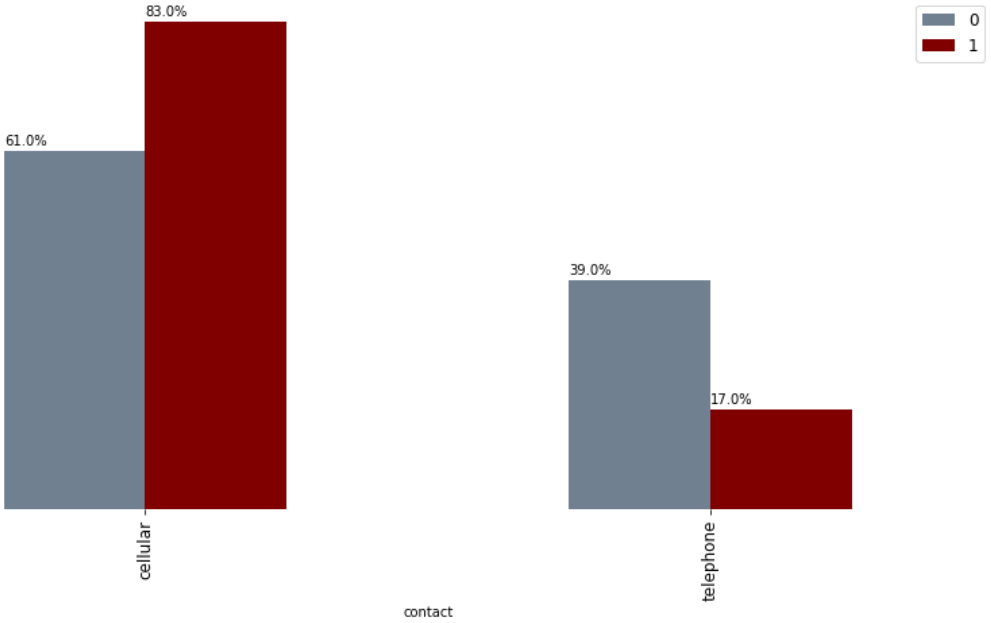


Tỉ lệ chuyển đổi theo các ngày trong tuần cũng tương đối đồng đều, cao nhất là thứ 5 với 12.11% và thấp nhất là thứ 2 với 9.95%. Khách hàng có xu hướng đăng ký tiền gửi có kỳ hạn nhiều hơn là từ chối vào thứ 3, thứ 4 và thứ 5. Tuy nhiên không có sự khác biệt đáng kể về số lượng khách hàng tiếp cận và số lượng người đăng ký nên có vẻ như không có ngày nào tốt hơn để liên hệ với những người đăng ký tiềm năng.

- **What type of contacts should we contact our customers?**



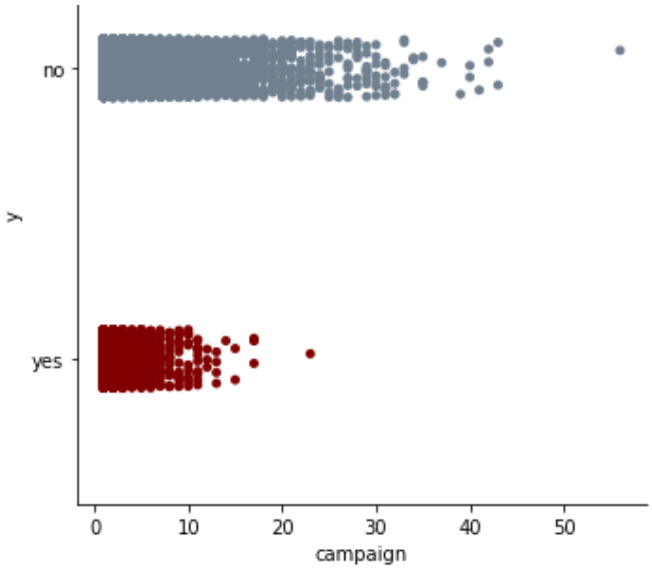
Với chiến dịch này, ngân hàng sử dụng 2 phương thức là ‘cellular’ và ‘telephone’ để liên hệ với khách hàng. Hình thức liên lạc qua ‘cellular’ chiếm 63.47% và có tỉ lệ chuyển đổi cao gấp gần 3 lần so với ‘telephone’. Có lẽ do thời đại công nghệ phát triển mạnh mẽ nên mọi người sử dụng điện thoại di động nhiều hơn, điện thoại bàn chỉ để ở nhà và khả năng cao được sử dụng bởi những người lớn tuổi. Chính vì vậy, khi nhìn vào biểu đồ dưới đây, chúng ta có thể thấy rõ rằng tỉ lệ khách hàng liên hệ qua ‘cellular’ đồng ý cũng nhiều hơn so với khách hàng qua ‘telephone’ gấp 5 lần.



Action: Do khách hàng sử dụng điện thoại di động có nhiều khả năng đăng ký gửi tiết kiệm hơn nhóm khách hàng sử dụng điện thoại bàn nên ở chiến dịch tiếp theo, ngân hàng nên tập trung gọi điện tiếp thị cho nhóm khách hàng sử dụng thuê bao di động.

- **How many times should we contact the customer?**

Theo mô tả dữ liệu, campaign là chỉ số chỉ ra số lần khách hàng được liên hệ trong chiến dịch này.

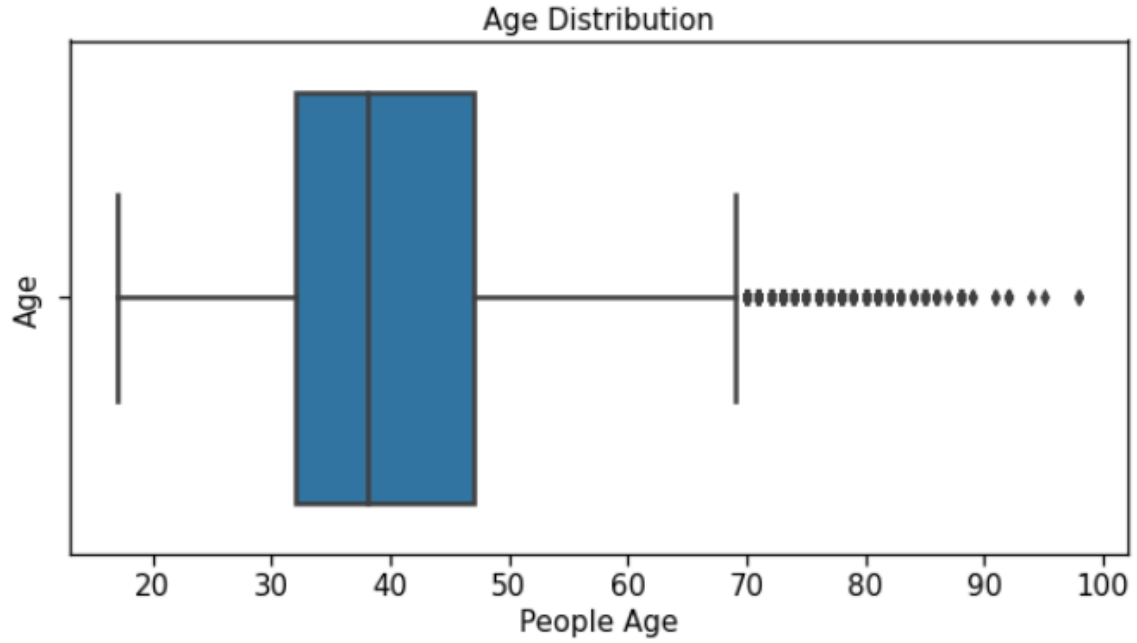


Khách hàng hầu như chấp nhận mở tài khoản trước lần liên hệ thứ 10. Trong các trường hợp, khách hàng đa số đăng ký trước lần liên hệ thứ 6.
=> Liên lạc với một khách hàng quá 10 lần dường như không hữu ích.
Action: Nếu liên hệ với khách hàng tới lần thứ 10 mà họ không đồng ý đăng kí thì bỏ qua khách hàng đó để tiết kiệm chi phí, thời gian, nhân công.

Phần 2: Modeling

1. Visual some columns data

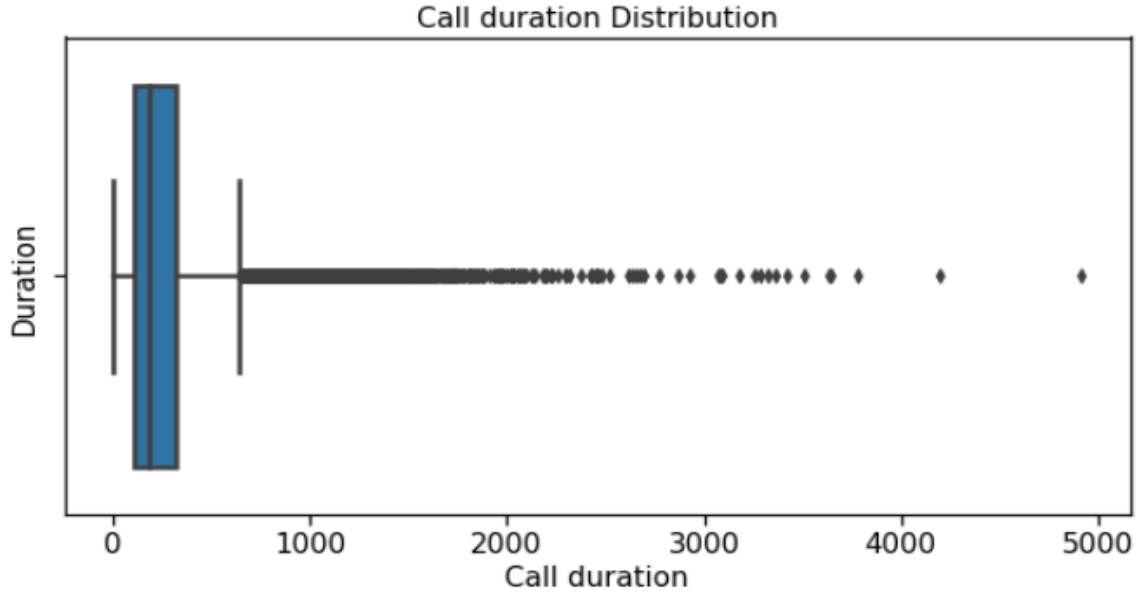
- **Age**



Ta có thể thấy outliner của Age trên 70 tuổi rất nhiều. Do đó, Age sẽ chia thành các nhóm để giảm bớt outliner.

- Dưới 32 tuổi : nhóm 1
- Trên 32 và dưới 47 tuổi : nhóm 2
- Trên 47 và dưới 70 tuổi : nhóm 3
- Trên 70 tuổi : nhóm 4

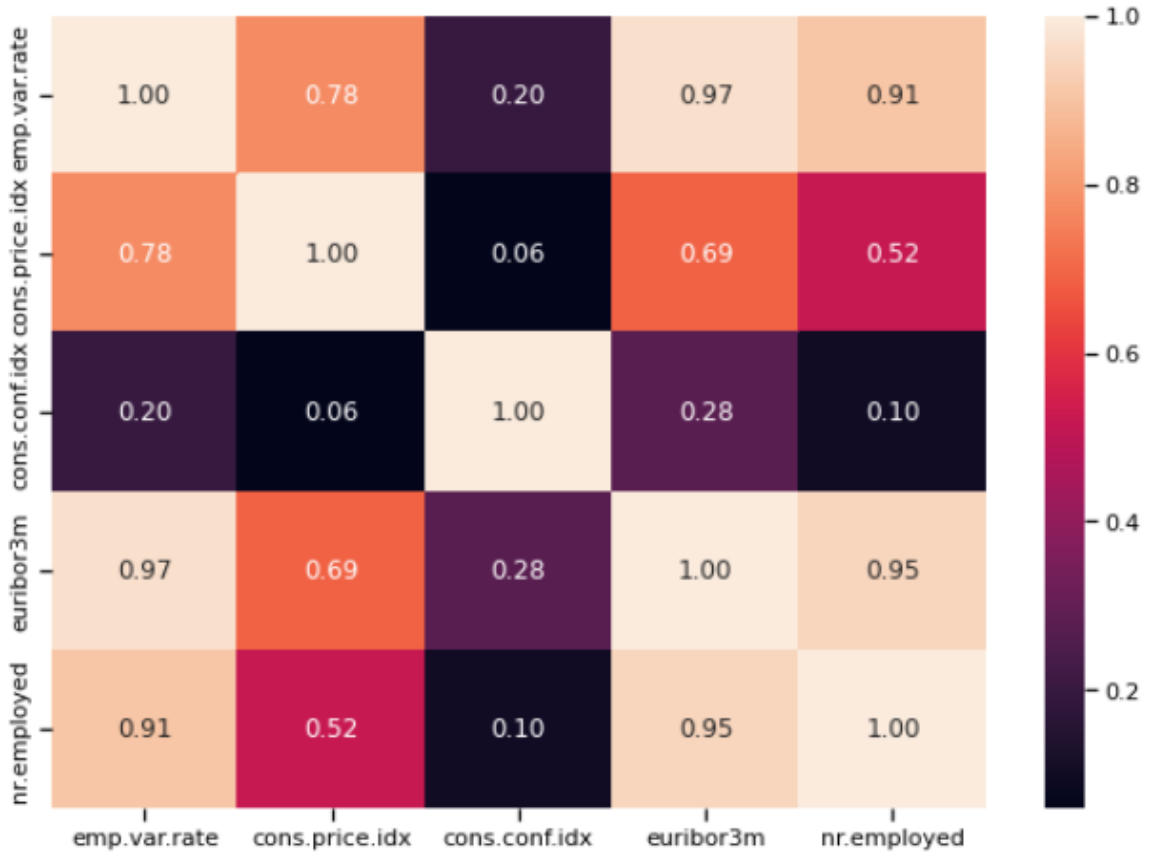
- **Duration**



Tương tự như Age, Duration cũng có nhiều outliner, do đó chúng ta cũng chia khoảng.

- Duration <= 102 : 1
- 102 < Duration <= 180 : 2
- 180 < Duration <= 319 : 3
- 319 < Duration <= 645 : 4
- Duration > 645 : 5

- **Corr**



euribor3m variable có corr cao với 3 biến là emp.var.rate, cons.price.idx & nr.employed. Do đó chúng ta sẽ xóa 3 biến này.

2. Processing

```
def Encoder(df):
    le = LabelEncoder()
    df = pd.get_dummies(df, columns = ['job','marital', 'education','default', 'loan', 'housing'])
    df['contact'] = df['contact'].map({'telephone': 1, 'cellular': 0})
    df['poutcome'] = df['poutcome'].map({'success': 1, 'nonexistent': 0, 'failure':0})
    df['y'] = df['y'].map({'yes':1, 'no':0})
    return df
```

Map cột contact và poutcome,
One-hot cho các cột job, marital, education, default, loan, housing.
Replace giá trị yes, no cột y.

```
def clean(df):
    df.dropna(inplace=True)
    df = df.drop(columns=['emp.var.rate','cons.price.idx', 'nr.employed', 'month', 'day_of_week'])
    df = df.drop_duplicates()
    return df
```

Drop các giá trị duplicated và một số cột như emp.var.rate, cons.price.idx, nr.employed, month, day_of_week.

3. Train-test

```
def train_test(X, y):
  X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
  k_fold = KFold(n_splits=10, shuffle=True, random_state=0)
  sc_X = StandardScaler()
  X_train = sc_X.fit_transform(X_train)
  X_test = sc_X.transform(X_test)
  return X_train, X_test, y_train, y_test
```

- Chia tập train và test với 7:3
- Sử dụng kFold, sau đó scaler X_train và X_test.

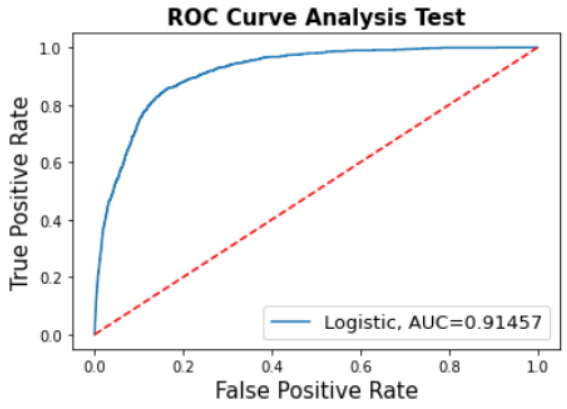
4. Chọn model

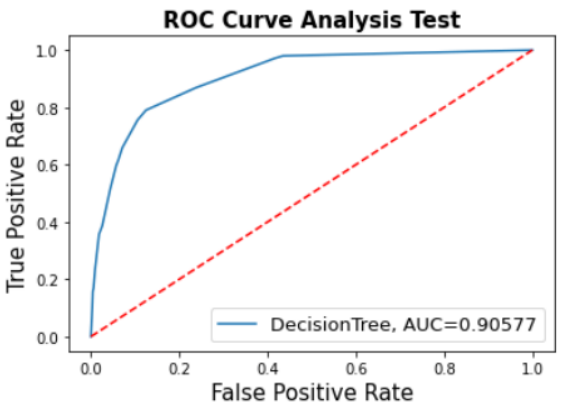
```
Logistic
recall_score: 0.38660907127429806
f1_score: 0.4988388295401765
precision_score: 0.7028795811518325


DecisionTree
recall_score: 0.4895608351331893
f1_score: 0.4881550610193826
precision_score: 0.4867573371510379
AUC train
```

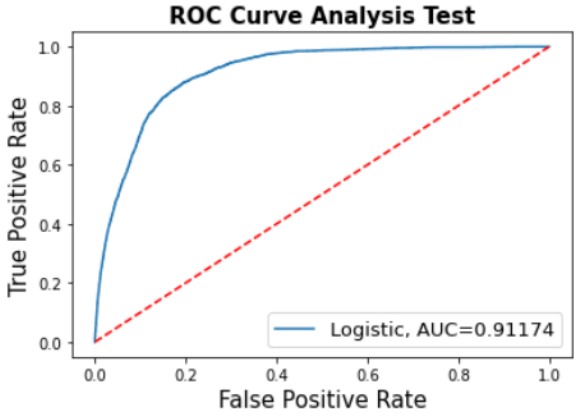
Sau khi Hyperparameter Tuning.

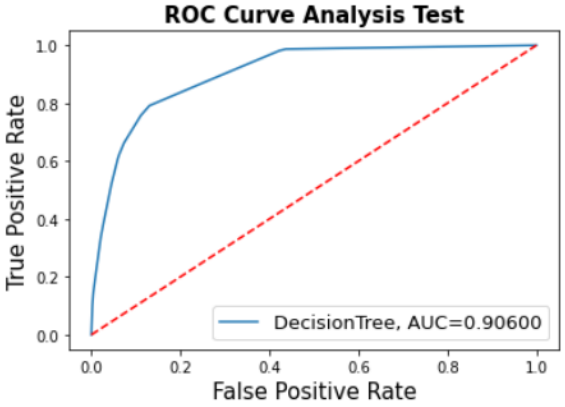
AUC test





AUC train





Ta có thể thấy ở mode Logistic, chỉ số AUC ở cả tập train và test gần như không có sự chênh lệch quá lớn như model Decision Tree.
Chọn model Logistic.