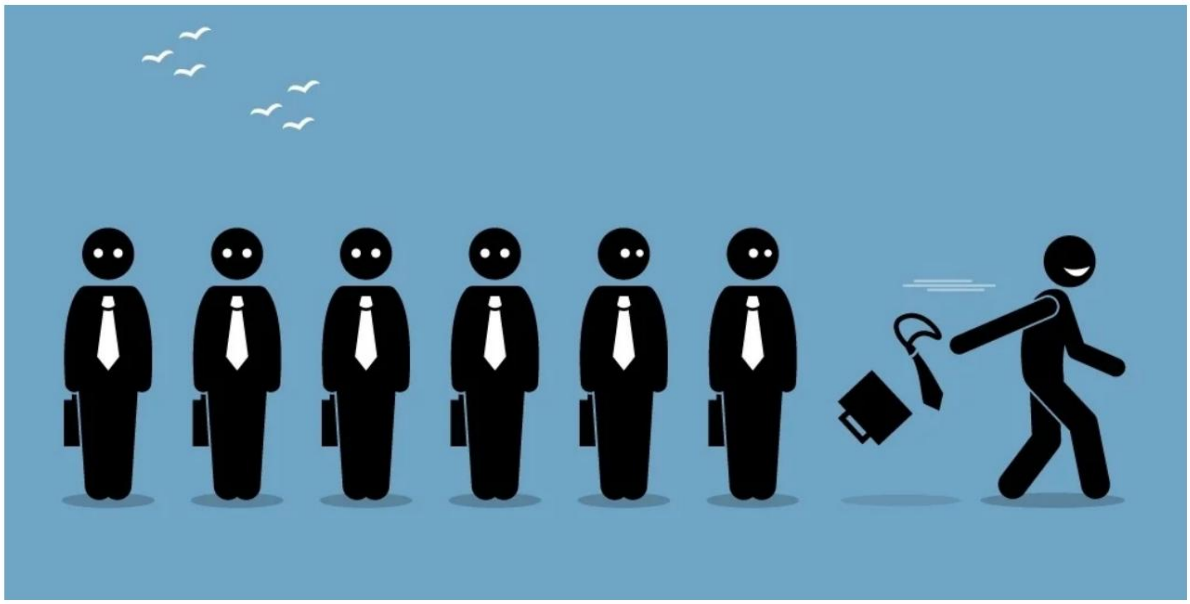
Dự án có hướng dẫn: Làm sạch và phân tích nhân viên Thoát khảo sát

Trong dự án có hướng dẫn này, chúng tôi sẽ làm việc với các cuộc khảo sát về việc thôi việc từ các nhân viên của Bộ Giáo dục, Đào tạo và Việc làm (DETE) và viện Giáo dục Kỹ thuật và Nâng cao (TAFE) ở Queensland, Australia. Bạn có thể tìm khảo sát về việc rời khỏi TAFE tại đây (https://data.gov.au/dataset/ds-qld-89970a3b-182b-41ea-aea2-6f9f17b5907e/details?q=exit%20survey) và khảo sát cho DETE tại đây (https://data.gov.au/dataset/ds-qld-fe96ff30-d157-4a81-851d-215f2a0fe26d/details?q=exit%20survey).



Trong dự án này, chúng tôi sẽ cố gắng trả lời các câu hỏi sau:

- Có phải những nhân viên chỉ làm việc cho các viện trong một thời gian ngắn đã từ chức vì một số loại không hài lòng? Còn những nhân viên đã ở đó lâu hơn thì sao?
- Có phải nhân viên trẻ tuổi từ chức vì một số loại không hài lòng? Còn những nhân viên lớn tuổi thì sao?

Chúng tôi sẽ kết hợp kết quả của cả hai cuộc khảo sát để trả lời những câu hỏi này. Bạn có thể tìm thấy mô tả về các cột của bộ dữ liệu trong tệp README.

Hãy bắt đầu bằng cách đọc các bộ dữ liệu và khám phá chúng.

khám phá dữ liệu

Trong [134]:
```
nhập gấu trúc dưới dạng
pd nhập numpy dưới dạng np
```

Trong [135]:
```
dete_survey = pd.read_csv("dete_survey.csv")
tafe_survey = pd.read_csv("tafe_survey.csv")
```

Trong [136]:
```
dete_survey.info()
```

```
<lớp 'pandas.core.frame.DataFrame'>
Range Index: 822 mục, 0 đến 821
Các cột dữ liệu (tổng cộng 56 cột):
 # Cột                                 Dtype đếm không null
--- ------                                 -------------- -----
 0 ID                                  822 không rỗng           int64
 1 Loại phân tách                      822 không rỗng 822       sự vật
 2 Ngày Ngừng                          không rỗng 822           sự vật
 3 PHÁT HIỆN Ngày bắt đầu              không rỗng 822           sự vật
 4 Ngày bắt đầu vai trò                không rỗng               sự vật
 5 vị trí                              817 không rỗng 455       sự vật
 6 Phân loại                           không rỗng 822           sự vật
 7 khu vực                             không rỗng 126           sự vật
 8 Đơ n vị kinh doanh                  không rỗng               sự vật
 9 Tình trạng việc làm                 817 không rỗng 822       sự vật
 10 Sự nghiệp chuyển sang khu vực công không rỗng 822           bool
 11 Sự nghiệp chuyển sang khu vực tư  nhân không rỗng 822       bool
 12 Xung đột giữa các cá nhân          không rỗng               bool
 13 Sự không hài lòng trong công việc 822 non-null              bool
 14 Không hài lòng với bộ phận 822 non-null                     bool
 15 Môi trư ờng làm việc thể chất 822 non-null                  bool
 16 Thiếu sự công nhận 822 non-null                             bool
 17 Thiếu bảo đảm việc làm             822 không rỗng 822       bool
 18 Địa điểm làm việc                  không rỗng 822           bool
 19 Điều kiện tuyển dụng               không rỗng 822           bool
 20 Sản phụ/gia đình                   không rỗng               bool
 21 Di dời                             822 không rỗng 822       bool
 22 Du Học/Du Lịch                     không rỗng 822           bool
 23 Bệnh tật                           không rỗng 822           bool
 24 Tai nạn thư ơ ng tâm               không rỗng               bool
 25 Cân bằng cuộc sống công việc       822 không rỗng 822       đối
 26 Khối lư ợng công việc              không rỗng 822           tư ợng
 27 Không có ý nào ở trên              không rỗng 808           bool
 28 Phát Triển Nghề Nghiệp             không rỗng               bool bool
 29 Cơ  hội thăng tiến                 735 không rỗng 816       sự vật
 30 Tinh thần nhân viên                không rỗng 788           sự vật
 31 Vấn đề nơ i làm việc               không rỗng 817           sự vật
 32 Môi trư ờng vật chất               không rỗng               sự vật
 33 Cân bằng cuộc sống công việc       815 không rỗng 810       sự vật
 34 Hỗ trợ căng thẳng và áp lực        không rỗng 813           sự vật
 35 Công việc của ngư ời giám sát      không rỗng 812           sự vật
 36 Hỗ trợ đồng đẳng                   không rỗng               sự vật
 37 Sáng kiến                          813 không rỗng 811       sự vật
 38 kỹ năng                            không rỗng 767           sự vật
 39 huấn luyện viên                    không rỗng 746           sự vật
 40 nguyện vọng nghề nghiệp            không rỗng               sự vật
 41 Phản hồi                           792 không rỗng 768       sự vật
 42 Tiếp theo PD                       không rỗng 814           sự vật
 43 Giao tiếp                          không rỗng 812           sự vật
 44 tôi nói                            không rỗng               sự vật
 45 Thông tin                          816 không rỗng 813       sự vật
 46 Đư ợc thông báo                    không rỗng 766           sự vật
 47 chư ơ ng trình chăm sóc sức khỏe   không rỗng 793           sự vật
 48 Sức khỏe & An toàn                 không rỗng               sự vật
 49 giới tính                          798 không rỗng 811       sự vật
 50 tuổi                               không rỗng 16            sự vật
 51 thổ dân                            không rỗng               sự vật
```

```
52 eo biển Torres                    3 không rỗng          sự vật
53 Biển Nam                          7 không rỗng 23       sự vật
54 Khuyết tật                        không rỗng 32         sự vật
55 NESB                              không rỗng            sự vật
dtypes: bool(18), int64(1), đối tư ợng(37)
sử dụng bộ nhớ: 258,6+ KB
```

Trong [137]:
```
tafe_survey.head()
```

Ra[137]:

	ID bản ghi	Khu vực làm việc của Viện	ĐÌNH CHỈ NĂM	Lý do chấm dứt thuê ngư ời làm	Đóng góp Các nhân tố. Sự nghiệp Di chuyển - Công cộng ngành	Đóng góp Các nhân tố Sự nghiệp Di chuyển Riêng tư ngành
0	6.341330e+17	Phía Nam Queensland viện nghiên cứu TAFE	không Vận chuyển (công ty)	2010.0	Hợp đồng Hết hạn	NaN	NaN
1	6.341337e+17	Núi Isa viện nghiên cứu TAFE	không Vận chuyển (công ty)	2010.0	Sự nghỉ hư u	-	-
		Núi Isa	Vận chuyển				

Trong [138]:
```
dete_survey ['SeparationType'].value_counts()
```

Hết[138]:
```
Tuổi hư u                                    285
Từ chức-Lý do khác                           150
Từ chức-Chủ khác                              91
Từ chức-Di chuyển ra nư ớc ngoài/liên bang    70
Nghỉ hư u sớm tự nguyện (VER)                 67
Nghỉ hư u ốm đau                              61
Khác                                          49
Hợp đồng hết hạn                              34
chấm dứt                                      15
Tên: Tách Loại, dtype: int64
```

Trong [139]:
```
tafe_survey.isnull().head()
```

Hết[139]:

	Ghi chú	Khu vực làm việc của Viện	ĐÌNH CHỈ NĂM	Lý do chấm dứt thuê ngư ời làm	Đóng góp Các nhân tố. Sự nghiệp Di chuyển - Công cộng ngành	Đóng góp Các nhân tố. Sự nghiệp Di chuyển - Riêng tư ngành	đóng góp Nhân tố cẩn thận Di chuyển - Se công nhân	
0	SAI	SAI	SAI	SAI	SAI	ĐÚNG VẬY	ĐÚNG VẬY	Trù
1	SAI	SAI	SAI	SAI	SAI	SAI	SAI	Sai
2	SAI	SAI	SAI	SAI	SAI	SAI	SAI	Sai
3	sai	SAI	SAI	SAI	SAI	SAI	SAI	sai
4	Sai	SAI	SAI	SAI	SAI	SAI	SAI	Sai

5 hàng × 72 cột

Làm sạch dữ liệu

Từ một số ô trên, chúng ta thấy:

- Khung dữ liệu dete_survey chứa các giá trị 'Không đư ợc nêu' cho biết các giá trị là bị thiếu, như ng chúng không đư ợc biểu diễn dư ới dạng NaN.
- Cả khung dữ liệu dete_survey và tafe_survey đều chứa nhiều cột mà chúng tôi không cần phải hoàn thành phân tích của chúng tôi.
- Mỗi khung dữ liệu chứa nhiều cột giống nhau, như ng tên cột khác nhau.
- Có nhiều cột/câu trả lời cho biết một nhân viên đã nghỉ việc vì họ không hài lòng.

Để bắt đầu, chúng ta sẽ xử lý hai vấn đề đầu tiên.

Trong [140]:
```
#Đầu tiên, hãy mở lại bộ dữ liệu dete_survey
#như ng lần này thay thế các giá trị 'Không đư ợc nêu' bằng NaN

dete_survey = pd.read_csv("dete_survey.csv", na_values='Not Stated') dete_survey.head(3)
```

Hết[140]:

	Loại phân tách ID	Ngư ng Ngày	phát hiện Bắt đầu Ngày	Vai trò Bắt đầu Ngày	Khu vực phân loại vị trí			Việc kinh doanh Đơ n vị	em
0	1	Sức khỏe kém Sự nghỉ hư u	08/2012	1984,0 2004,0	Công cộng Ngư ời hầu	A01-A04	Trung tâm Văn phòng	công ty Chiến lư ợc và Hiệu suất	P
1	2	Tự nguyện sớm Sự nghỉ hư u (VER)	08/2012	NaN NaN	Công cộng Ngư ời hầu	AO5-AO7	Trung tâm Văn phòng	công ty Chiến lư ợc và Hiệu suất	P
2	3	Tự nguyện sớm Sự nghỉ hư u (VER)	05/2012	2011.0 2011.0	trư ờng học Nhân viên văn phòng	NaN	Trung tâm Văn phòng	Giáo dục Queensland	P

3 hàng × 56 cột

cột DROP

Trong [141]:
```
#Hãy bỏ các cột mà chúng tôi sẽ không sử dụng từ cuộc khảo sát DETE:
dete_survey_updated = dete_survey.drop(dete_survey.columns[28:49],axis = 1)
```

Mã này có thể đư ợc sử dụng:

```
column_to_drop = dete_survey.iloc[:,28:49] dete_survey_updated =
dete_survey.drop(columns_to_drop, axis=1)

# Hãy loại bỏ các cột mà chúng tôi sẽ không sử dụng từ cuộc khảo sát TAFE:
```

Trong [142]:
```
#Hãy bỏ các cột mà chúng tôi sẽ không sử dụng từ khảo sát TAFE:
tafe_survey_updated = tafe_survey.drop(tafe_survey.columns[17:66],axis = 1)
```

Ở trên, chúng tôi đã thực hiện trư ớc một số bư ớc làm sạch dữ liệu:

- 'Không đư ợc nêu' trong cuộc khảo sát dete đã đư ợc thay thế bằng NaN.
- Các cột sẽ không đư ợc sử dụng đã bị loại bỏ. Các phiên bản cập nhật của bộ dữ liệu đư ợc gán cho hai khung dữ liệu mới - dete_survey_updated và tafe_survey_updated .

Đổi tên các cột

Tiếp theo, hãy chú ý đến các tên cột. Mỗi khung dữ liệu chứa nhiều cột giống nhau, như ng tên cột khác nhau.

Trong [143]:
```
#Ví dụ: Ngày ngừng phải đư ợc cập nhật thành ngày ngừng
dete_survey_updated.columns = dete_survey_updated.columns.str.replace(" ","_")
```

Trong [144]:
```
dete_survey_updated.head(3)
```

Ra[144]:

	loại	phân tách	id	ngừng_ngày	dete_start_date	vai trò_bắt đầu_ngày	phân loại	vị trí	regio
0	1	Sức khỏe kém Sự nghỉ hư u	08/2012	1984.0	2004.0	Công cộng Ngư ời hầu	A01-A04	trung tâm chính thức	
1	2	Tự nguyện sớm Sự nghỉ hư u (VER)	08/2012	NaN	NaN	Công cộng Ngư ời hầu	A05-A07	trung tâm chính thức	
2	3	Tự nguyện sớm Sự nghỉ hư u (VER)	05/2012	2011.0	2011.0	trư ờng học Nhân viên văn phòng	NaN	trung tâm chính thức	

3 hàng × 35 cột

Trong [145]:
```
ánh xạ = {'ID bản ghi': 'id',
 'NGƯ ỜI NGỪNG': 'ngày_ngừng',
 'Lý do thôi việc': 'separationtype',
 'Giới tính. Giới tính của bạn là gì?': 'giới tính',
 'Tuổi hiện tại. Tuổi hiện tại': 'tuổi',
 'Loại việc làm. Loại Việc làm': 'job_status',
 'Phân loại. Phân loại': 'vị trí',
 'LengthofServiceOverall. Tổng thời gian phục vụ tại Viện (tính theo năm)': '

tafe_survey_updated = tafe_survey_updated.rename(ánh xạ, trục = 1)
```

Trong [146]:
```
tafe_survey_updated.head(3)
```

Ra[146]:

	~~	Loại phân tách	WorkArea	ngừng_ngày	của Viện		Đóng góp Các nhân tố. Sự nghiệp Di chuyển - Công cộng ngành	Đóng góp Các nhân tố. Sự nghiệp Di chuyển - Riêng tư ngành
0	6.341330e+17	Phía Nam Queensland viện nghiên cứu TAFE	không Vận chuyển (công ty)	2010.0	Hợp đồng Hết hạn	NaN	NaN	
1	6.341337e+17	Núi Isa viện nghiên cứu TAFE	không Vận chuyển (công ty)	2010.0	Sự nghỉ hư u	-	-	
2	6.341388e+17	Núi Isa viện nghiên cứu TAFE	Vận chuyển (giảng bài)	2010.0	Sự nghỉ hư u	-	-	

3 hàng × 23 cột

Trong một vài ô ở trên, tên của các cột trong cả hai khung dữ liệu đã đư ợc cập nhật.

Lọc dữ liệu

Tiếp theo, hãy xóa thêm dữ liệu mà chúng tôi không cần. Mục tiêu cuối cùng của chúng tôi là trả lời những điều sau đây
câu hỏi:

- Là nhân viên chỉ làm việc cho các viện trong một thời gian ngắn xin nghỉ việc
  do một số loại không hài lòng? Còn những nhân viên đã từng làm việc thì sao?
  lâu hơ n?

Điều này có nghĩa là chúng tôi chỉ quan tâm đến những nhân viên đã nghỉ việc:

Trong [147]:
```
dete_survey_updated['separationtype'].value_counts()
```

Hết[147]:
```
Tuổi hư u                                    285
Từ chức-Lý do khác                           150
Từ chức-Chủ khác                              91
Từ chức-Di chuyển ra nư ớc ngoài/liên bang    70
Nghỉ hư u sớm tự nguyện (VER)                 67
Nghỉ hư u ốm đau                              61
Khác                                          49
Hợp đồng hết hạn                              34
chấm dứt                                      15
Tên: tách loại, dtype: int64
```

Trong [148]:
```
tafe_survey_updated['separationtype'].value_counts()
```

Hết[148]:
```
Từ chức                340
Hợp đồng hết hạn       127
Cắt giảm/Dự phòng      104
Sự nghỉ hư u            82
Chuyển khoản            25
chấm dứt                23
Tên: tách loại, dtype: int64
```

Tại [149]:
```
# Sửa tên cột trong DETE, bỏ dấu "-" và chỉ để lại từ đầu tiên trong cột tên
# Hãy chú ý đến thứ tự của các phư ơ ng thức chuỗi đư ợc vector hóa, bởi vì chúng ta không'
dete_survey_updated['separationtype'] = dete_survey_updated['separationtype'].
```

Trong [150]:
```
dete_survey_updated['separationtype'].value_counts()
```

Hết[150]:
```
Từ chức                311
Tuổi về hư u           285
Nghỉ hư u sớm tự nguyện (VER)    67
Nghỉ hư u ốm đau                 61
Khác                             49
Hợp đồng hết hạn                 34
chấm dứt                         15
Tên: tách loại, dtype: int64
```

Trong [151]:
```
#Lọc dữ liệu
dete_resignations = dete_survey_updated[dete_survey_updated["separationtype"]
```

Trong [152]:
```
tafe_resignations = tafe_survey_updated[tafe_survey_updated['separationtype']
```

Ở trên, hai khung dữ liệu mới đã được tạo - dete_resignations và tafe_resignations. Họ giữ dữ liệu chỉ dành cho những nhân viên đã nghỉ việc.

xác minh dữ liệu

Tiếp theo, chúng tôi sẽ tập trung vào việc xác minh rằng các năm trong ngày thôi_ngày và ngày_bắt_đầu cột có ý nghĩa.

- Vì ngày thôi việc là năm cuối cùng làm việc của người đó và dete_start_date là năm đầu tiên làm việc của người đó, sẽ không hợp lý nếu có nhiều năm sau ngày hiện tại.
- Cho rằng hầu hết mọi người trong lĩnh vực này bắt đầu làm việc ở độ tuổi 20, cũng không chắc rằng dete_start_date là trước năm 1940.

Nếu chúng ta có nhiều năm cao hơn ngày hiện tại hoặc thấp hơn năm 1940, chúng ta sẽ không muốn tiếp tục với phân tích của chúng tôi, bởi vì nó có thể có nghĩa là có điều gì đó rất sai với dữ liệu. Nếu có một lượng nhỏ giá trị cao hoặc thấp phi thực tế, chúng tôi có thể xóa chúng.

Trong [153]:
```
dete_resignations["cease_date"].value_counts()
```

Hết[153]:
```
2012            126
2013             74
01/2014          22
12/2013          17
06/2013          14
09/2013          11
07/2013           9
11/2013           9
10/2013           6
08/2013           4
05/2012           2
05/2013           2
07/2006           1
09/2010           1
07/2012           1
2010              1
Tên: ngừng_date, dtype: int64
```

Trong [154]:
```
# Lọc năm
dete_resignations["cease_date"] = dete_resignations["cease_date"].str.split("/
```

Trong [155]:
```
dete_resignations["cease_date"] = dete_resignations["cease_date"].astype(float
```

Trong [156]:
```
dete_resignations['cease_date'].value_counts().sort_index()
```

Hết[156]:
```
2006.0      1
2010.0      2
2012.0    129
2013.0    146
2014.0     22
Tên: ngừng_date, dtype: int64
```

Trong [157]:
```
dete_resignations['dete_start_date'].value_counts().sort_index(ascending=True)
```

Ra[157]:
```
1963.0      1
1971.0      1
1972.0      1
1973.0      1
1974.0      2
1975.0      1
1976.0      2
1977.0      1
1980.0      5
1982.0      1
1983.0      2
1984.0      1
1985.0      3
1986.0      3
1987.0      1
1988.0      4
1989.0      4
1990.0      5
1991.0      4
1992 0      6
```

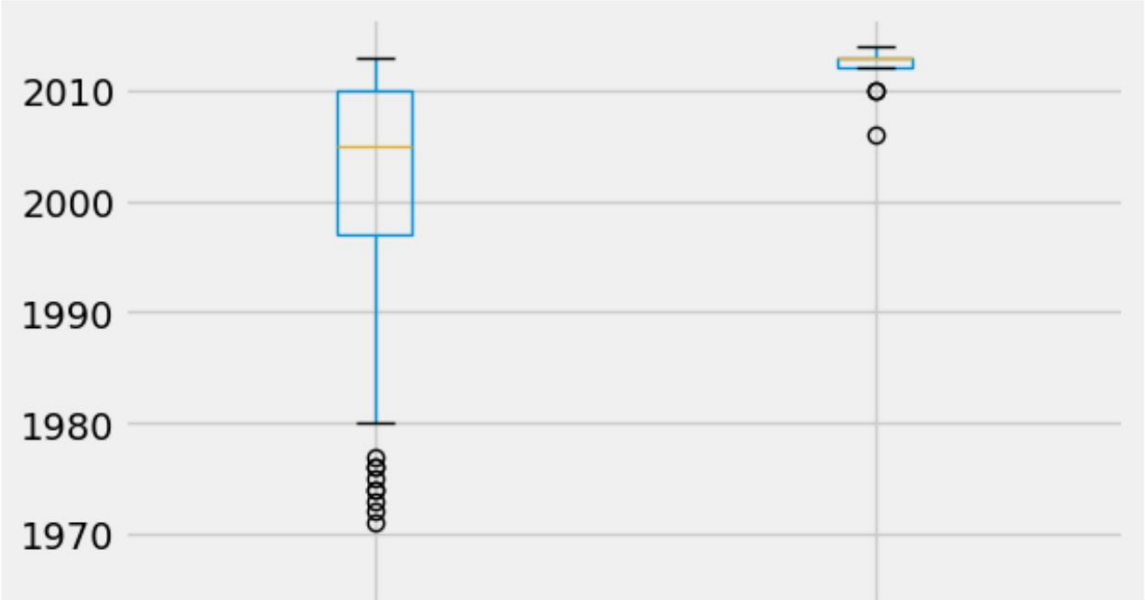
Trong [158]:
```
tafe_resignations['cease_date'].astype(float).value_counts().sort_index()
```

Hết[158]:
```
2009.0      2
2010.0     68
2011.0    116
2012.0     94
2013.0     55
Tên: ngừng_date, dtype: int64
```

Trong [159]:
```
nhập matplotlib.pyplot dư ới dạng plt
%matplotlib inline
plt.style.use(' fivethirtyeight')

dete_resignations[['dete_start_date','cease_date']].plot(kind='box')
```

Ra[159]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f245c31b580>



Ở trên, chúng tôi đã làm sạch các cột của cả hai khung dữ liệu chứa ngày bắt đầu và ngày kết thúc của các nhân viên đã thôi việc. Dư ờng như không có bất kỳ vấn đề lớn nào với các giá trị. Khoảng thời gian ngừng hoạt động của cả hai khung dữ liệu hơ i khác một chút:

- THÁM: 2006 - 2014
- TAFE: 2009 - 2013

Trong lĩnh vực Nhân sự, khoảng thời gian một nhân viên làm việc tại nơ i làm việc đư ợc gọi là số năm làm việc của họ.

Bộ dữ liệu TAFE chứa một cột có tên là viện_dịch vụ . Thật không may, bộ dữ liệu DETE không có cột như vậy. Tuy nhiên, chúng tôi có dữ liệu cần thiết để tạo cột này. Nó phải chứa sự khác biệt giữa các cột ngừng_date và dete_start_date .

Trong [160]:
```
dete_resignations['instA_service'] = dete_resignations['cease_date'] - det
```

Xác định nhân viên không hài lòng

Tiếp theo, chúng tôi sẽ xác định bất kỳ nhân viên nào đã từ chức vì họ không hài lòng. Dư ới đây là các cột chúng tôi sẽ sử dụng để phân loại nhân viên là "không hài lòng" từ mỗi khung dữ liệu:

1. TAFE:
    - Yếu tố góp phần. Các yếu tố góp phần
    - gây ra sự không hài lòng. Bất mãn với công việc
2. PHÁT HIỆN:

- Sự không hài lòng công việc
- không hài lòng_với_bộ_phận
- thể chất_công việc_môi trư ờng
- Thiếu sự công nhận
- thiếu_of_job_security
- trụ sở làm việc
- việc làm_điều kiện
- cân bằng cuộc sống công việc
- khối lư ợng công việc

Nếu nhân viên chỉ ra bất kỳ yếu tố nào ở trên khiến họ từ chức, chúng tôi sẽ đánh dấu họ là không hài lòng trong một cột mới.

Chúng tôi sẽ bắt đầu với dữ liệu TAFE.

Trong [161]:
```
tafe_resignations[' Các yếu tố đóng góp. Không hài lòng'].value_counts(dropna
```

Hết[161]:
```
- Yếu tố góp phần. Tên NaN không hài lòng: Các yếu tố đóng góp.
Không hài lòng, dtype: int64
```
277
55

Trong [162]:
```
tafe_resignations[' Các yếu tố đóng góp. Sự không hài lòng về công việc '].value_counts(dr
```

Ra[162]:
```
-                                   270
Bất mãn với công việc                62
Tên: Các yếu tố đóng góp. Không hài lòng về công việc, dtype: int64
```

Chúng ta cần chuyển đổi những phản hồi này thành giá trị boolean và NaN. Vì vậy, bây giờ chúng ta sẽ tạo một chức năng để làm điều tư ơ ng tự.

Trong [163]:
```
def update_vals(x):
    nếu pd.isnull(x):
        trả về np.nan
    yêu tinh x == '-':
        trả về Sai
    khác:
        trả về Đúng
```

Trong [164]:
```
tafe_cols = ['Các yếu tố đóng góp. Sự không hài lòng',' Các yếu tố góp phần. Jo

#Cập nhật cột tafe_resignations với bool
tafe_resignations[tafe_cols] = tafe_resignations[tafe_cols].applymap(update_va
```

Trong [165]:
```
tafe_resignations[' Các yếu tố đóng góp. Không hài lòng'].value_counts(dropna
```

Hết[165]:
```
Sai          277
ĐÚNG VẬY      55
Tên: Các yếu tố đóng góp. Không hài lòng, dtype: int64
```

Trong [166]:
```
tafe_resignations[' Các yếu tố đóng góp. Sự không hài lòng về công việc '].value_counts(dr
```

Hết[166]:
```
Sai 270
ĐÚNG VẬY 62
Tên
NaN: Các yếu tố đóng góp. Không hài lòng về công việc, dtype: int64
```

Bây giờ chúng ta có thể thấy rằng chúng ta đã chuyển đổi thành công các giá trị thành Bool. Bây giờ chúng ta có thể đi phía trư ớc và áp dụng logic của chúng tôi để tìm ra những ngư ời không hài lòng. Bất cứ ai trả lời đúng sẽ đư ợc đánh dấu là không hài lòng.

Trong [167]:
```
# Tạo cột mới có tên 'không hài lòng' để lư u trữ giá trị

tafe_resignations['không hài lòng'] = tafe_resignations[tafe_cols].any(axis=1, s
```

Trong [168]:
```
tafe_resignations['không hài lòng'].value_counts(dropna=False)
```

Hết[168]:
```
Sai 241
ĐÚNG VẬY 91
NaN
Tên: không hài lòng, dtype: int64
```

Chúng ta có thể thấy rằng cột mới đã đư ợc tạo và nó đang lư u trữ các giá trị boolean bất mãn .

Không hài lòng với bộ dữ liệu DETE

Bây giờ chúng tôi cũng thực hiện các bư ớc tư ơ ng tự với bộ dữ liệu DETE. Để làm đư ợc điều đó chúng ta sẽ phải tìm chỉ mục của các cột để dễ dàng thao tác.

Trong [169]:
```
dete_diss_columns = ['job_dissatisfaction', 'dissatisfaction_with_the_departme
 'lack_of_recognition', 'lack_of_job_security', 'work_location', 'job_co
 'work_life_balance', 'khối lư ợng công việc']

dete_resignations['không hài lòng'] = dete_resignations[dete_diss_columns].any(a
```

Trong [170]:
```
dete_resignations['không hài lòng'].value_counts(dropna=False)
```

Hết[170]:
```
Sai 162
ĐÚNG VẬY 149
Tên: không hài lòng, dtype: int64
```

Trong [171]:
```
dete_resignations_up = dete_resignations.copy()
tafe_resignations_up = tafe_resignations.copy()
```

Ở trên, chúng tôi đã tạo một cột không hài lòng trong cả hai khung dữ liệu. Giá trị của các cột là Đúng hoặc Sai dựa trên câu trả lời của nhân viên đối với các câu hỏi trong các cột mà chúng tôi xác định ở trên.

Ngoài ra, chúng tôi đã tạo các bản sao của từng khung dữ liệu.

Kết hợp dữ liệu

Bây giờ chúng tôi đã sẵn sàng kết hợp hai bộ dữ liệu mà chúng tôi đang làm việc thành một. Chúng tôi sẽ tổng hợp tập dữ liệu của mình dựa trên cột Institute_service.

Thêm định danh Viện

Để dễ dàng xác định các hàng sau khi tổng hợp, chúng tôi sẽ thêm một cột viện cho cả hai bộ dữ liệu.

Trong [172]:
```
dete_resignations_up["viện"] = "DETE"
tafe_resignations_up["viện"] = "TAFE"
## Kết hợp
kết hợp = pd.concat([dete_resignations_up, tafe_resignations_up], ignore_inde
```

Trong [173]:
```
## Kiểm tra dữ liệu bị thiếu
kết hợp.notnull().sum().sort_values(ascending=False)
```

Ra[173]: id 651

học viện
651
tách loại 651

bất mãn
643
ngừng_ngày
635
vị trí 598

việc làm_tình trạng
597

tuổi 596
giới
tính 592
viện_dịch vụ 563

Khu vực làm
việc 340
học viện
340
Yếu tố góp phần. Học 332

Yếu tố góp phần. KHÔNG CÓ 332

Yếu tố góp phần. Xung đột giữa các cá nhân 332

Yếu tố góp phần. Khác 332

Yếu tố góp phần. Không hài lòng 332

Yếu tố góp phần. Bư ớc Chuyển Sự Nghiệp - Khu Vực Công 332

Yếu tố góp phần. Chuyển nghề - Khu vực tư nhân 332

Yếu tố góp phần. Di Chuyển Sự Nghiệp - Tự Kinh Doanh 332

Yếu tố góp phần. Bệnh tật 332

Yếu tố góp phần. Thai sản/Gia đình 332

Yếu tố góp phần. Du lịch 332

Yếu tố góp phần. Không hài lòng với công việc 332

công việc_không hài lòng
311
Thiếu sự công nhận
311
vật_lý_làm_việc_môi-trư ờng 311

bất mãn_với_bộ_phận 311

trụ sở làm việc

311

liên nhân_xung đột 311

Career_move_to_private_sector 311

Career_move_to_public_sector 311

thiếu_việc_làm_bảo_vệ 311

khối lượng công việc

311

việc làm_điều kiện 311

thai sản/gia đình 311

di dời 311

du học/du lịch
311

bệnh_sức_khỏe
311

chấn thương_sự cố 311

work_life_balance 311

none_of_the_above
311

Độ dài dịch vụ hiện tại. Thời gian làm việc tại nơi làm việc hiện tại (tính theo năm) 290

dete_start_date 283

role_start_date
271

vùng 265

phân loại 161

doanh nghiệp_đơn
vị 32

nesb

9

khuyết tật 8

thổ dân
7

nam_biển 3

torres_strait 0

dtype: int64

Chúng ta có thể thấy rằng hầu hết các cột mà chúng ta cần phân tích thêm đều có hơn 500 giá trị khác null. Vì vậy, chúng tôi có thể đặt 500 làm ngưỡng để loại bỏ các giá trị khác null.

Trong [174]:
```
##Lọc dữ liệu
kết hợp_updated = kết hợp.dropna(thresh = 500, trục =1).copy()
tổ hợp_updated.head()
```

Hết[174]:

	loại	tách	id	ngừng_ngày	vị trí	việc làm_tình trạng	giới tính	tuổi	học viện_dịch vụ
0	4.0	Sự từ chức		2012.0	Giáo viên	Cố định Toàn thời gian	Nữ	36-40	7
1	6,0	Sự từ chức		2012.0	hướng dẫn Nhân viên văn phòng	Cố định Toàn thời gian	Nữ	41-45	18
2	9,0	Sự từ chức		2012.0	Giáo viên	Cố định Toàn thời gian	Nữ	31-35	3
3	10,0	Sự từ chức		2012.0	Giáo viên phụ tá	Nữ Bán Thời Gian Cố Định		46-50	15
4	12,0	Sự từ chức		2012.0	Giáo viên	Thường trực Toàn thời gian	Nam giới	31-35	3

Trong một vài ô ở trên, chúng tôi đã làm như sau:

- đã tạo một cột viện trong mỗi tập dữ liệu cho biết nơi nhân viên làm việc;
- kết hợp hai bộ dữ liệu thành một bộ dữ liệu mới được gọi là kết hợp;
- đã xóa bất kỳ cột nào khỏi tập dữ liệu mới có hơn 500 giá trị NaN. Kết quả tập dữ liệu đã được chỉ định cho tổ hợp_updated.

Làm sạch cột 'viện_dịch vụ'

Tiếp theo, chúng ta cần xóa cột Institute_service vì nó chứa các giá trị trong một vài các định dạng khác nhau. Để phân tích dữ liệu, chúng tôi sẽ chuyển đổi những con số này thành các danh mục. Chúng tôi sẽ căn cứ phân tích của chúng tôi về bài viết này (https://www.businesswire.com/news/home/20171108006002/en/Age-Number-Engage-Employee-Career-Stage), đưa ra lập luận rằng việc hiểu nhu cầu của nhân viên theo giai đoạn nghề nghiệp thay vì tuổi hiệu quả hơn.

Chúng tôi sẽ sử dụng các định nghĩa được sửa đổi một chút bên dưới:

- Mới: Dưới 3 năm làm việc tại công ty
- Kinh nghiệm: 3-6 năm tại công ty
- Thành lập: 7-10 năm tại công ty
- Cựu chiến binh: 11 năm trở lên tại công ty

Trong [175]:
```
tổ hợp_updated['institute_service'].value_counts()
```

Hết[175]:

Dư ới 1 năm 1-2	73
	64
3-4	63
5-6	33
20-11	26
5.0	23
1.0	22
7-10	21
3.0	20
0,0	20
6,0	17
4,0	16
9,0	14
2.0	14
7,0	13
Hơ n 20 năm 13,0	10
8,0	
20,0	7
15,0	7
14,0	6
17,0	6
12,0	6
10,0	6
22,0	6
18,0	5
16,0	5
24,0	4
23,0	4
11,0	4
39,0	3
19,0	3
21,0	3
32,0	3
25,0	2
26,0	2
36,0	2
28,0	2
30,0	2
42,0	1
49,0	1
35,0	1
34,0	1
38,0	1
33,0	1
29,0	1
27,0	1
41,0	1
31,0	1

Tên: viện_service, dtype: int64

Trong [176]:
```
tổ hợp_updated["instA_service"].unique()
```

Ra[176]:
```
mảng([7.0, 18.0, 3.0, 15.0, 14.0, 5.0, nan, 30.0, 32.0, 39.0, 17.0, 9.0,
       6.0, 1.0, 35.0, 38.0, 36.0, 19.0, 4.0, 26.0, 10.0, 8.0, 2.0, 0.0, 23.0, 13.0, 16.0,
       12.0, 21.0, 20.0, 24.0, 33.0, 22.0, 2 8.0, 49.0, 11.0, 41.0, 27.0, 42.0, 25.0, 29.0,
       34.0, 31.0, '3-4', '7-10', '1-2', 'Dưới 1 năm', '11-20', '5-6', ' Hơn 20 năm'],

      dtype=đối tượng)
```

Trong [177]:
```
print(combined_updated["institute_service"].value_counts().sum())
```

```
563
```

Chúng ta có thể thấy rằng có hai loại dữ liệu trong cột này, một là số trong khi loại kia là phạm vi năm. Chúng ta có thể tiếp tục và phân loại chúng thành các nhóm.

Chúng tôi sẽ làm theo định nghĩa được đề cập dưới đây để nhóm.

- Mới: Dưới 3 năm làm việc tại công ty
- Kinh nghiệm: 3-6 năm tại công ty
- Thành lập: 7-10 năm tại công ty
- Cựu chiến binh: 11 năm trở lên tại công ty

Trước tiên, chúng tôi sẽ trích xuất số năm từ các giá trị này, sau đó so sánh và nhóm chúng thành các danh mục tương ứng.

Trong [178]:
```
## Trích xuất năm, sử dụng str.extract("(\d+)") để lọc các phạm vi, chúng ta có thể nhận được tổ
hợp_updated['institute_service_up'] = tổ hợp_updated['institute_service
print(combined_updated['institute_service_up'] .unique()) tổ
hợp_updated['inst a_service_up'] = tổ hợp_updated['instA_service ## Kiểm tra kết hợp_updated['inst

a_service_up'].value_counts().sum()
```

```
['7' '18' '3' '15' '14' '5' nan '30' '32' '39' '17' '9' '6' '1' '35' '38' '36' '19' '4' '26' '10'
 '8' '2' '0' '23' '13' '16' '12' '21' '20' '24' '33' '22' '28 ' '49' '11' '41' '27' '42' '25'
 '29' '34' '31']
```

Hết[178]:
```
563
```

Trong [179]:
```
tổ hợp_updated['inst a_service_up'].value_counts().head()
```

Hết[179]:
```
1,0 3,0        159
                83
5,0             56
7,0             34
11.0 30 Tên:
Institute_service_up, dtype: int64
```

Trong [180]:
```
# Phân loại năm làm việc thành các phân đoạn:
def transform_service(x):
    nếu pd.isnull(x):
        trả về np.nan
    eif x < 3:
        trả về 'Mới'
    eif 3 <= x < 7:
        trả về 'Có kinh nghiệm'
    eif 7 <= x < 11:
        trả về 'Đã thành lập'
    khác:
        trở lại 'Cựu chiến binh'
```

Trong [181]:
```
tổ hợp_updated['service_cat'] = tổ hợp_updated['institute_service_up'].app
```

Trong [182]:
```
#Kiểm tra
tổ hợp_updated['service_cat'].value_counts(dropna=False)
```

Hết[182]:
```
Mới                 193
Có kinh nghiệm      172
cựu chiến binh      136
NaN                  88
Thành lập            62
Tên: service_cat, dtype: int64
```

Ở trên, chúng tôi đã xóa cột Institute_service. Chúng tôi đã sử dụng giá trị từ cột đó trong để xác định nhân viên thuộc loại nào. Chúng tôi đã tạo một cột mới - service_cat - nơ i chúng tôi thấy danh mục của nhân viên.

Không hài lòng theo danh mục

Trong [183]:
```
tổ hợp_updated['không hài lòng'].value_counts(dropna=False)
```

Hết[183]:
```
Sai            403
ĐÚNG VẬY       240
NaN
Tên: không hài lòng, dtype: int64
```

Điền giá trị Null

Chúng ta có thể thấy rằng vẫn còn 8 hàng có giá trị NaN trong cột hài lòng. Vì đây là một tỷ lệ nhỏ các giá trị bị thiếu, chúng ta có thể thay thế các giá trị bị thiếu bằng hầu hết giá trị thư ờng xuyên, đó là Sai.

Trong [184]:
```
kết hợp_updated['không hài lòng'] = kết hợp_updated['không hài lòng'].fillna(Sai
```

Trong [185]:
```
tổ hợp_updated['không hài lòng'].value_counts(dropna=False)
```

Hết[185]:
```
Sai 411
ĐÚNG VẬY 240
Tên: không hài lòng, dtype: int64
```

Trực tiếp hóa mối quan hệ của tuổi với độ không hài lòng với công việc

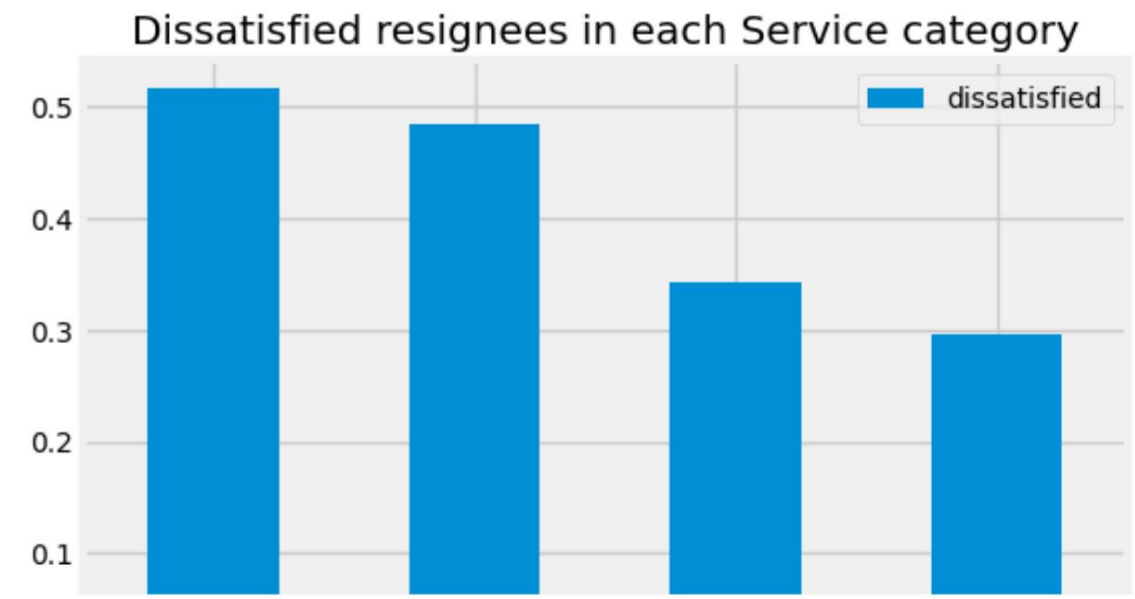
Trong [191]:
```
# bảng tần suất
#pd.pivot_table(combined_updated, index='service_cat', values='không hài lòng')
dis_pct = kết hợp_updated.pivot_table(index='service_cat', values='dissatisfi
dis_pct
```

Hết[191]:

dịch vụ_cat	bất mãn
Thành lập	0,516129
Có kinh nghiệm	0,343023
Mới	0,295337
cựu chiến binh	0,485294

Trong [197]:
```
# Vẽ sơ đồ nhân viên bất mãn trong từng loại dịch vụ
%matplotlib nội tuyến
dis_pct.sort_values('không hài lòng', tăng dần=Sai).plot(
    kind='bar', rot=30, title='Những ngư ời nghỉ việc không hài lòng trong mỗi loại Dịch vụ
```

Ra[197]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f245c133c40>



Ở trên, chúng tôi đã tạo một bảng tổng hợp tính tỷ lệ phần trăm nhân viên không hài lòng cho từng loại dịch vụ. Sau đó, chúng tôi vẽ kết quả trên biểu đồ thanh.

Chúng ta có thể thấy điều đó của những nhân viên đã đảm nhận hai chức vụ, Thành lập và Cựu chiến binh nhân viên có nhiều khả năng từ chức vì không hài lòng. Nhân viên mới ít có khả năng làm vì thế.

Có bao nhiêu ngư ời trong mỗi giai đoạn nghề nghiệp đã từ chức vì một số loại

Trong [199]:
```
dis_count = pd.pivot_table(combined_updated, index='service_cat', values='dis
dis_count = dis_count.sort_values(by=['không hài lòng'])
dis_count = dis_count.rename(columns={'không hài lòng':'không hài lòng_count'})
dis_count
```

Hết[199]:

dịch vụ_cat	không hài lòng_count
Thành lập	32,0
Mới	57,0
Có kinh nghiệm	59,0
cựu chiến binh	66,0

Ở trên, chúng tôi thấy số lư ợng ngư ời trong từng danh mục dịch vụ đã rời đi do không hài lòng.

Không hài lòng về độ tuổi

Dư ới đây chúng tôi sẽ làm sạch cột tuổi bằng cách nhóm các nhân viên theo nhóm tuổi. Sau đó, chúng tôi sẽ trả lời câu hỏi:

- Có bao nhiêu ngư ời trong mỗi nhóm tuổi đã từ chức vì một số loại không hài lòng?

Để làm sạch dữ liệu ta sẽ chia các nhóm tuổi như sau:

- tuổi 20
- tuổi 30
- thập niên 40
- thập niên 50
- 60+

Trong [200]:
```
tổ hợp_updated['age'].value_counts(dropna=False)
```

Hết[200]:
```
51-55                71
NaN                  55
41-45                48
41 45                45
46-50                42
36-40                41
46 50                39
26-30                35
21 25                33
31 35                32
26 30                32
36 40                32
56 trở lên           29
31-35                29
21-25                29
56-60                26
61 trở lên           23
20 tuổi trở xuống 10
Tên: tuổi, dtype: int64
```

Trong [201]:
```
tổ hợp_updated['tuổi'] = tổ hợp_updated['tuổi'].astype('str')
```

Trong [202]:
```
độ tuổi xác định :
    nếu s[0] == '2':
        quay lại 'tuổi 20'
    yêu tinh s[0] == '3':
        quay lại '30s'
    yêu tinh s[0] == '4':
        quay lại '40s'
    yêu tinh s[0] == '5':
        trở lại '50s'
    yêu tinh s[0] == '6':
        trở lại 'thập niên 60'
    yêu tinh s == 'nan':
        trả về np.nan
```

Trong [203]:
```
kết hợp_updated['tuổi'] = kết hợp_updated['tuổi'].apply(tuổi)
```

Trong [204]:
```
tổ hợp_updated['age'].value_counts()
```

Ra[204]:
```
40s 20s        174
               139
tuổi 30        134
thập           126
niên 50 60      23
Tên: tuổi, dtype: int64
```

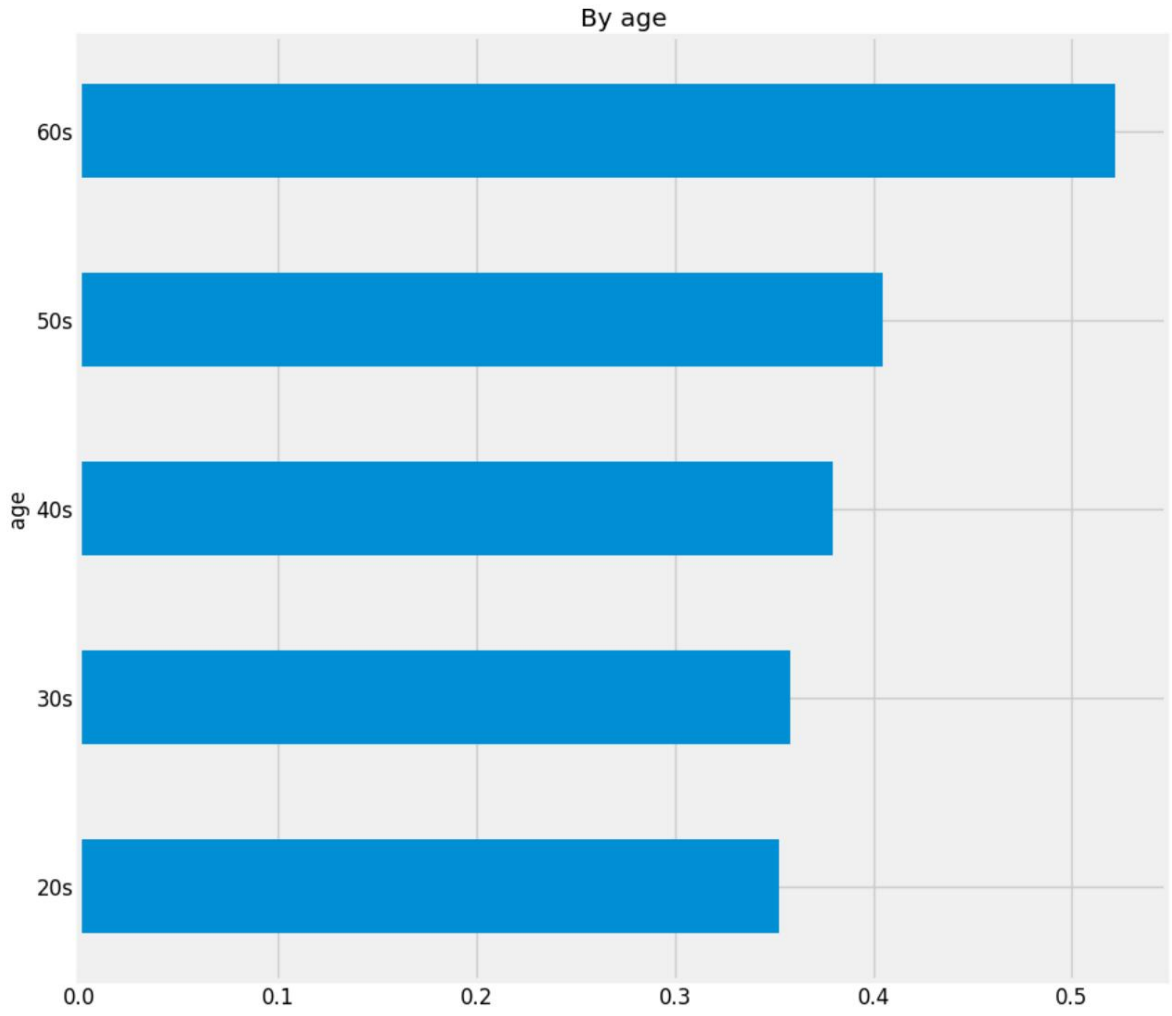
Trong [205]:
```
age_diss_count = pd.pivot_table(combined_updated, index='age', values='dissati
age_diss_count = age_diss_count.sort_values(by=['không hài lòng'])
age_diss_count = age_diss_count.rename(columns={'không hài lòng': 'không hài lòng_
age_diss_count
```

Hết[205]:

tuổi	không hài lòng_count
thập niên 60	12,0
tuổi 30	48,0
tuổi 20	49,0
thập niên 50	51,0
thập niên 40	66,0

Trong [206]:
```
age_perc = pd.pivot_table(combined_updated, index='age', values='dissatisfied'
age_perc = age_perc.sort_values(by=['không hài lòng'])
age_perc.plot(kind='barh', legend=False, figsize=(10,10), fontsize=12, title =
```

Ra[206]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457e43df0>



Trong các biểu đồ trên, chúng ta thấy:

1. Số ngư ời trong mỗi độ tuổi đã từ chức vì không hài lòng.
2. Tỷ lệ ngư ời trong mỗi nhóm tuổi đã từ chức vì không hài lòng.

Nhìn chung, số lư ợng nhân viên trẻ của hai viện xin nghỉ việc do không hài lòng với công việc là thấp nhất. Điều này có thể là do họ mới bắt đầu phát triển sự nghiệp và vẫn đang tìm kiếm một con đư ờng sự nghiệp để theo đuổi.

Không hài lòng với viện

- Có phải nhiều nhân viên hơ n trong cuộc khảo sát của DETE hoặc cuộc khảo sát của TAFE đã thôi việc vì họ không hài lòng theo một cách nào đó?

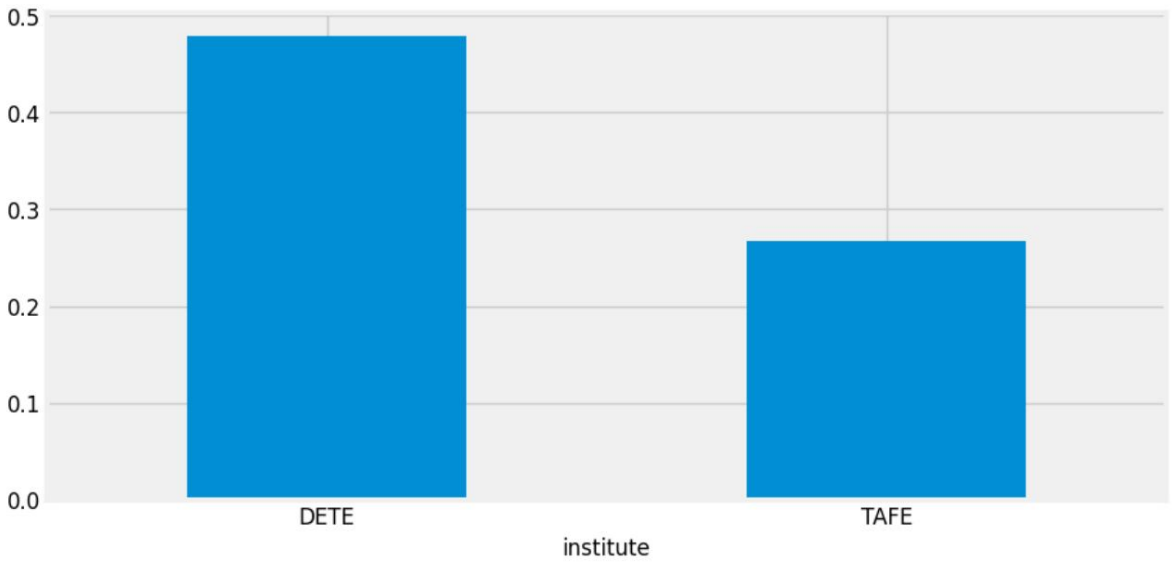
Trong [207]:
```
viện_count = pd.pivot_table(combined_updated, index='inst acad', values=' viện_count =
viện_count.rename(cột={'không hài lòng': 'không hài lòng viện_count
```

Hết[207]:

học viện	không hài lòng_count
phát hiện	149.0
TAFE	91,0

Trong [208]:
```
by_perc = pd.pivot_table(combined_updated, index='institute', values='dissatis
by_perc.plot(kind='bar', rot=360, figsize=(10,5), fontsize=12, legend =Sai)
```

Ra[208]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457f2f9a0>



Trong các biểu đồ trên, chúng ta thấy:

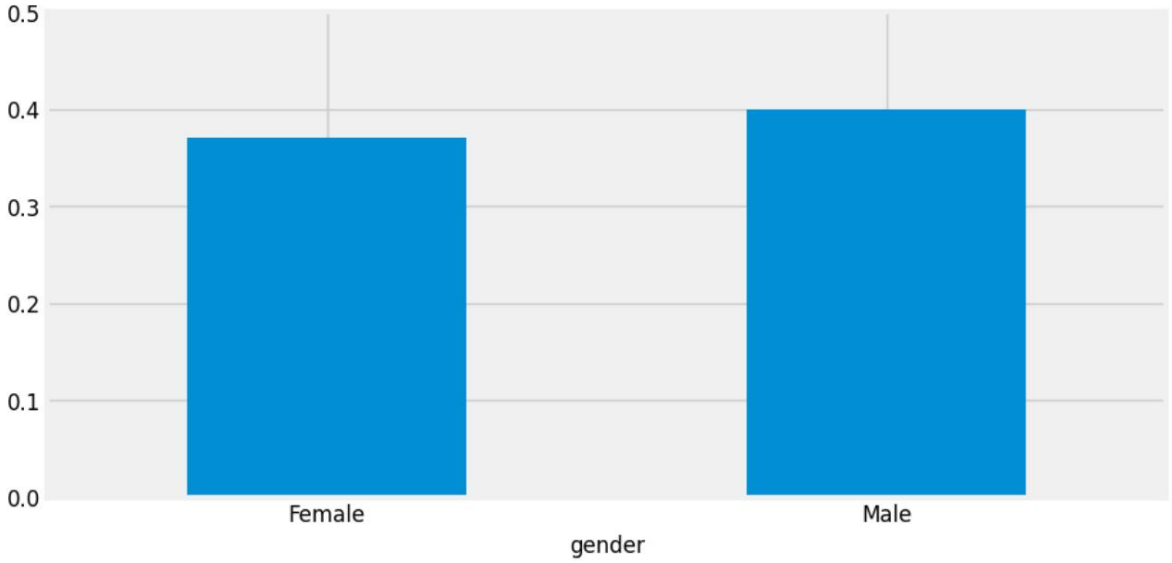
1. Số ngư ời từ mỗi viện đã từ chức vì không hài lòng.
2. Tỷ lệ ngư ời từ mỗi viện đã từ chức vì không hài lòng.

Có vẻ như nhân viên DETE đã từ chức vì không hài lòng với công việc thư ờng xuyên hơ n nhân viên TAFE.

Sự không hài lòng theo giới tính

Trong [209]:
```
by_gender = pd.pivot_table(combined_updated, index='gender', values='dissatisf
by_gender.plot(kind='bar', rot=360, figsize=(10,5), fontsize=12, ylim =[0, 0,5]
```

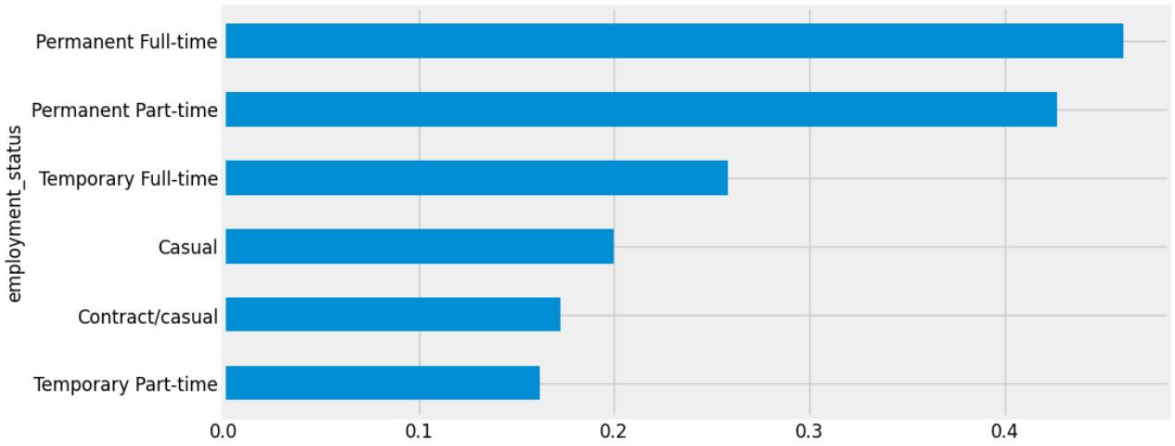
Ra[209]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457f7d1c0>



Không hài lòng với tình trạng việc làm

Trong [210]:
```
by_status = pd.pivot_table(combined_updated, index='employ_status', các giá trị
by_status = by_status.sort_values(by=['không hài lòng'])
by_status.plot(kind='barh', figsize=(10,5), fontsize=12, legend=False)
```

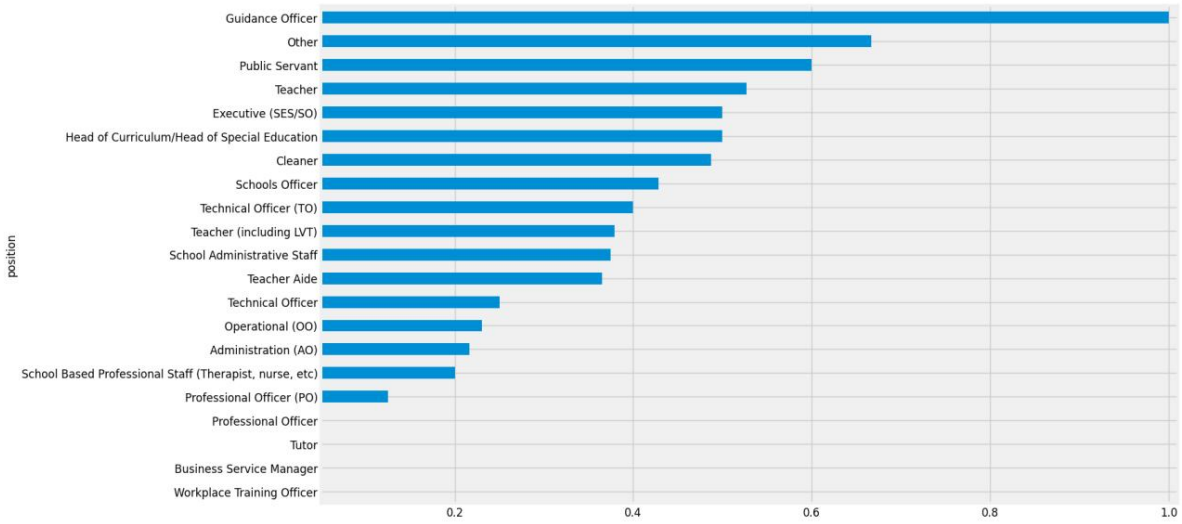
Ra[210]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457dc2eb0>



Không hài lòng với vị trí

Trong [211]:
```
by_position = pd.pivot_table(combined_updated, index='position', values='dissa
by_position = by_position.sort_values(by=['không hài lòng'])
by_position.plot(kind='barh', figsize=(15, 10), fontsize=12, legend=False, xli
```

Ra[211]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457f74a00>



Không hài lòng theo độ tuổi và giới tính

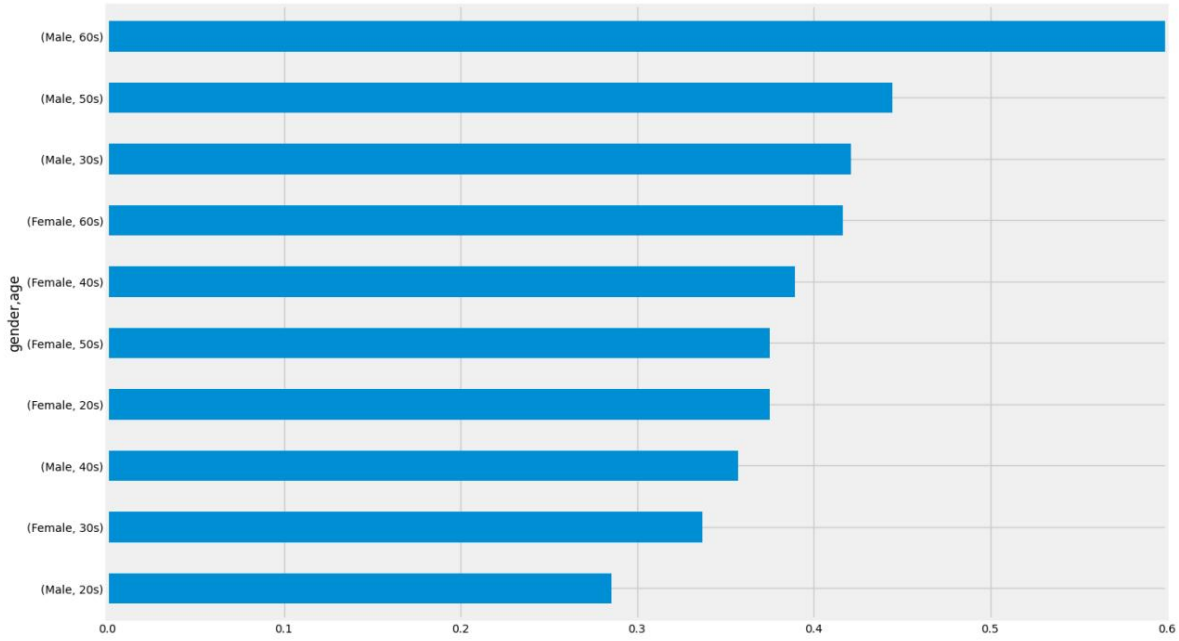
Trong [212]:
```
giới tính_tuổi = pd.pivot_table(combined_updated, index=['giới tính', 'tuổi'], giá trị=
giới_tuổi
```

Hết[212]:

		bất mãn
tuổi giới tính		
	tuổi 20	0,375000
	tuổi 30	0,336842
Nữ giới	thập niên 40	0,389313
	thập niên 50	0,375000
	thập niên 60	0,416667
	tuổi 20	0,285714
	tuổi 30	0,421053
Nam giới	thập niên 40	0,357143
	thập niên 50	0,444444
	thập niên 60	0,750000

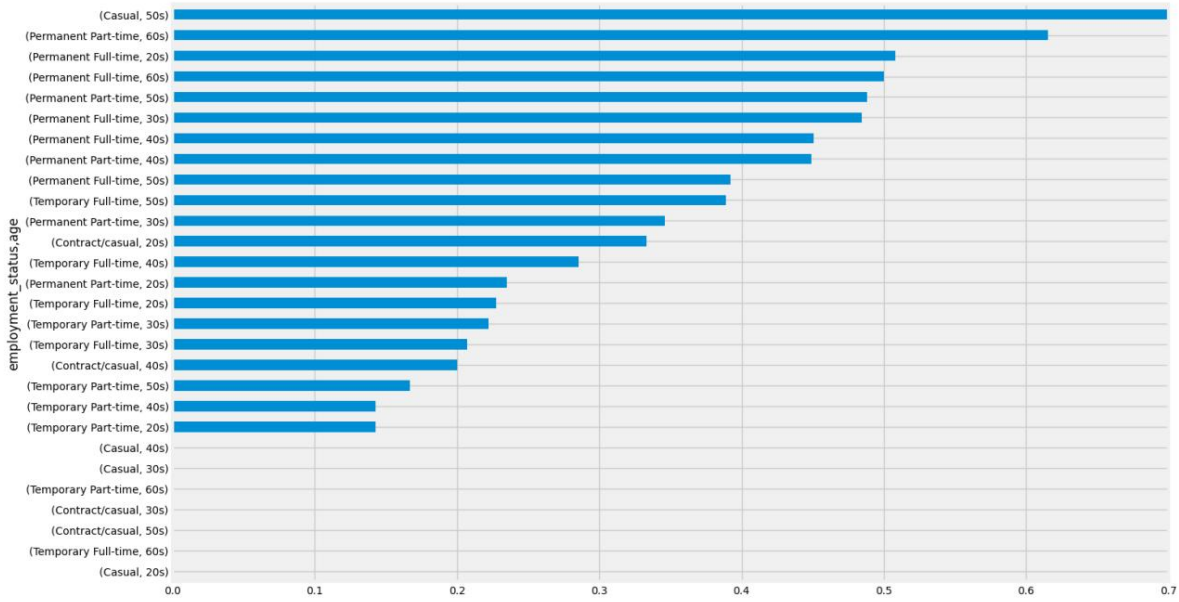
Trong [213]:
```
giới tính_tuổi = giới tính_tuổi.sort_values ( by= ['không
hài lòng'] )
```

Ra[213]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f245c21e850>



Trong [214]:
```
status_age = pd.pivot_table(combined_updated, index=['employ_status', 'age status_age
= status_age.sort_values(by=['không hài lòng'])
status_age.plot(kind='barh', figsize=(15 , 10), chú giải=Sai, xlim=[0, 0,7])
```

Ra[214]: <matplotlib.axes._subplots.AxesSubplot tại 0x7f2457e0af10>



Phần kết luận

Trong dự án này, chúng tôi đã phân tích các cuộc khảo sát về việc thôi việc của nhân viên của các viện DETE và TAFE. Chúng tôi tập trung vào những ngư ời đã từ chức do không hài lòng với công việc và kết luận rằng:

- nhân viên trẻ, thiếu kinh nghiệm ít từ chức nhất do không hài lòng với công
- việc; Nhân viên DETE từ chức thư ờng xuyên hơ n do không hài lòng với công việc so với
- nhân viên TAFE; Nam nhân viên nghỉ việc do không hài lòng với công việc nhiều hơ n Nữ
- nhân viên; Nam giới ở độ tuổi 20 ít khi từ chức vì không hài
- lòng nhất; 100% cán bộ Hư ớng dẫn đã từ chức đã điền vào các cuộc khảo sát đã từ chức do không hài lòng.