1. 데이터 파악 및 칼럼 탐구

무결한 데이터로 학습하기 위하여 칼럼에 대한 탐구와 그를 통한 오류치 발견에 집중하였다.

칼럼은 아래의 표로 구성되어 있다.

컬럼명	설명
age	나이 (숫자)
job	직업 (범주형)
marital	결혼 여부 (범주형)
education	교육 수준 (범주형)
default	신용 불량 여부 (범주형)
housing	주택 대출 여부 (범주형)
loan	개인 대출 여부 (범주형)
contact	연락 유형 (범주형)
month	마지막 연락 월 (범주형)
day_of_week	마지막 연락 요일 (범주형)
duration	마지막 연락 지속 시간, 초 단위 (숫자)
campaign	캠페인 동안 연락 횟수 (숫자)
pdays	이전 캠페인 후 지난 일수 (숫자)
previous	이전 캠페인 동안 연락 횟수 (숫자)
poutcome	이전 캠페인의 결과 (범주형)
emp.var.rate	고용 변동률 (숫자)
cons.price.idx	소비자 물가지수 (숫자)
cons.conf.idx	소비자 신뢰지수 (숫자)
euribor3m	3개월 유리보 금리 (숫자)
nr.employed	고용자 수 (숫자)
y	정기 예금 가입 여부 (**'y'** 또는 **'no'**로 표시됨)

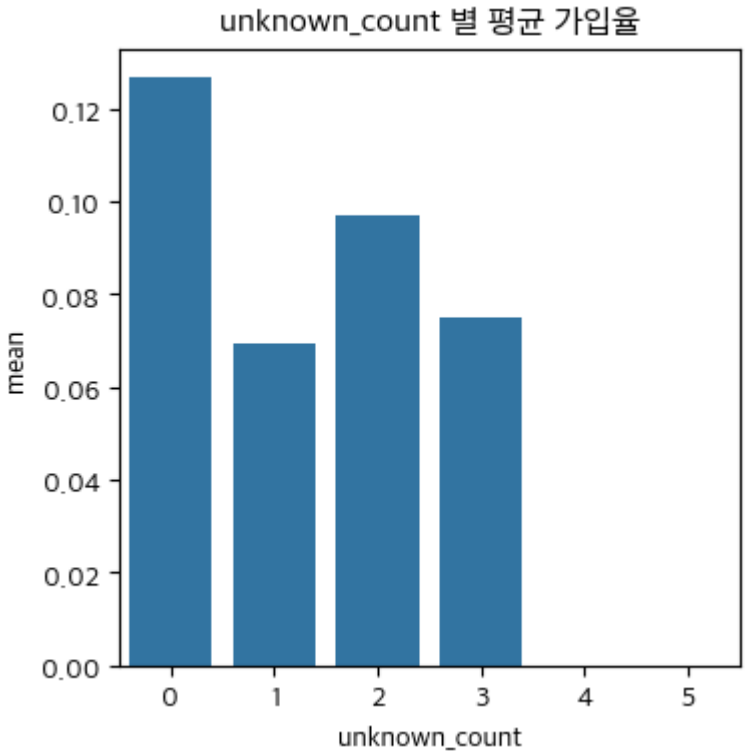
1-1. 고객 정보

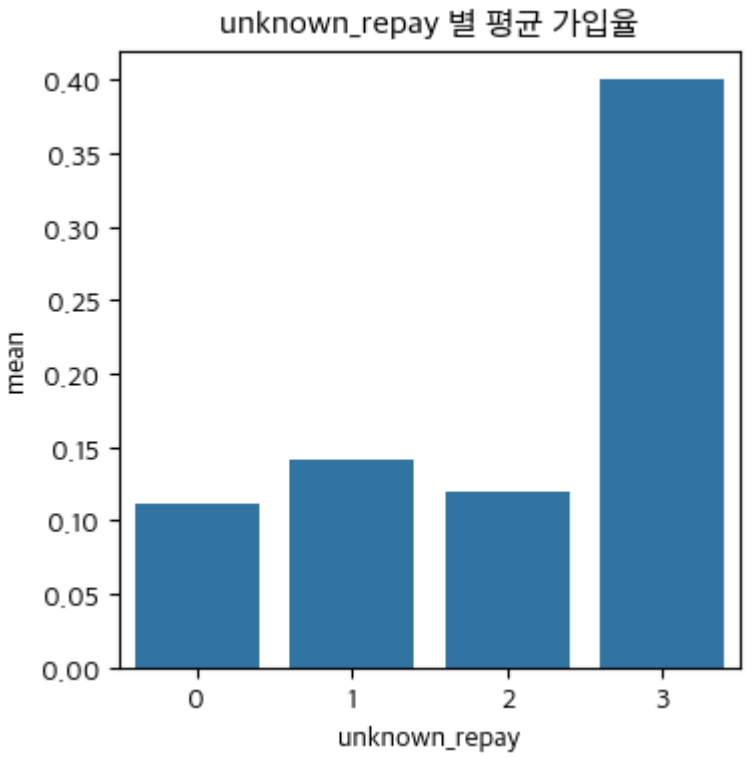
- 상환 능력 (`age` `job` `marital` `education`)
- 신용 정보 (`default` `housing` `loan`)

고객 정보는 상환 능력과 신용 정보로 이루어져 있다. 이 중 일부 항목의 응답값에는 'Unknown' 항목이 존재한다.

가설: Unknown 항목이 많을수록 가입률이 낮을 것이다.

- `unknown_repay` 와 `unknown_credit` 칼럼을 만들어, 상환 능력 관련 칼럼 중 Unknown 항목의 개수를 나타내도록 하였다.
- `unknown_count` 칼럼을 통해 총 개수를 표현하였다.
- 각 칼럼을 그룹으로 두어 가입률을 조사하였다.

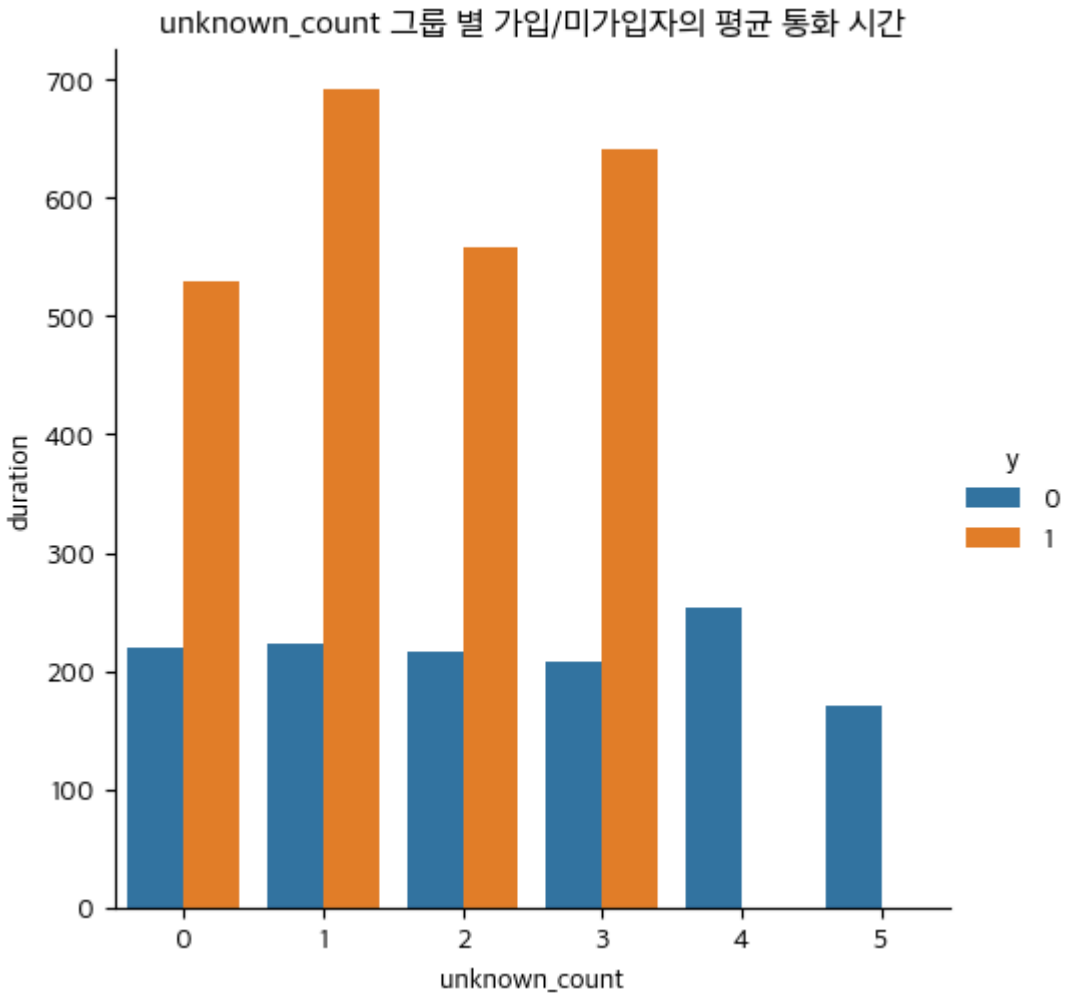

unknown_count 별 평균 가입율
mean
0.12
0.10
0.08
0.06
0.04
0.02
0.00
0
1
2
3
4
5
unknown_count

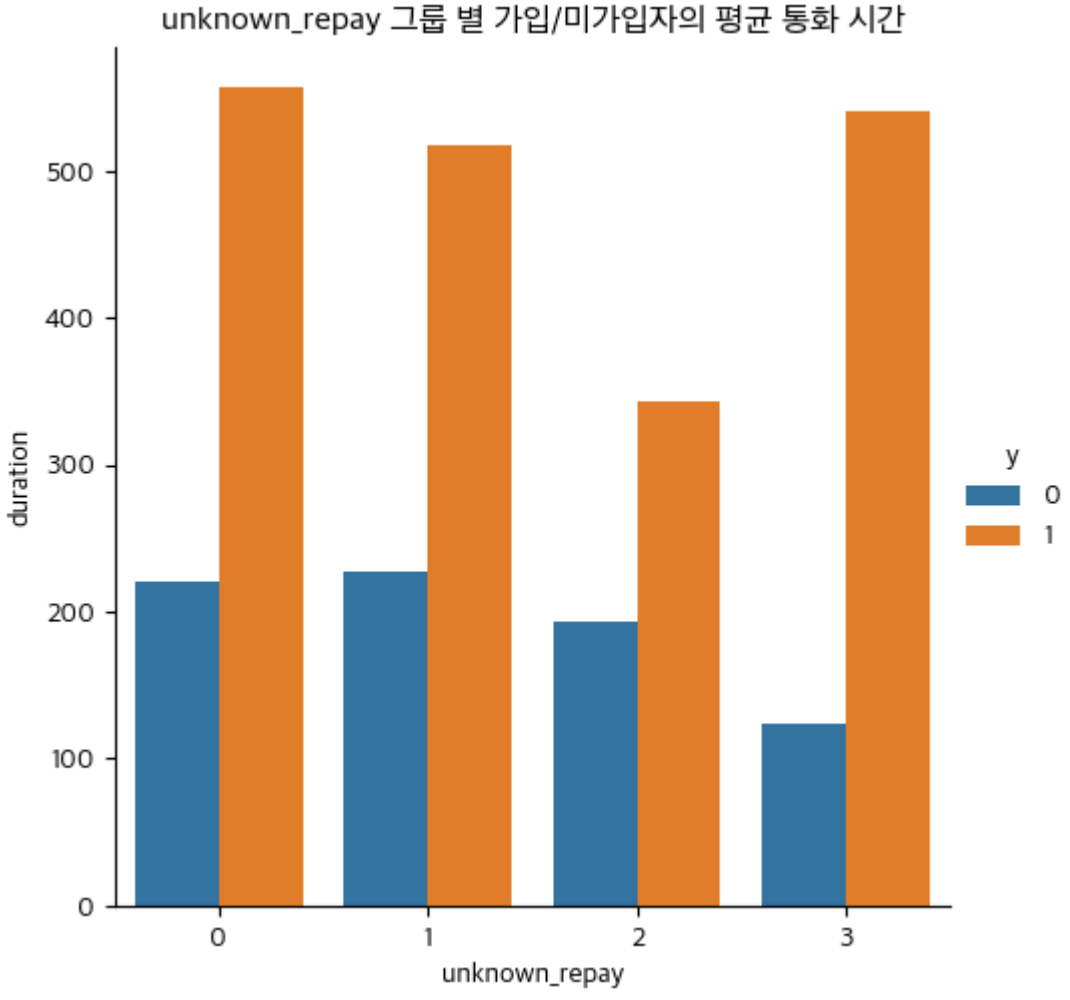

unknown_repay 별 평균 가입율
mean
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0
1
2
3
unknown_repay

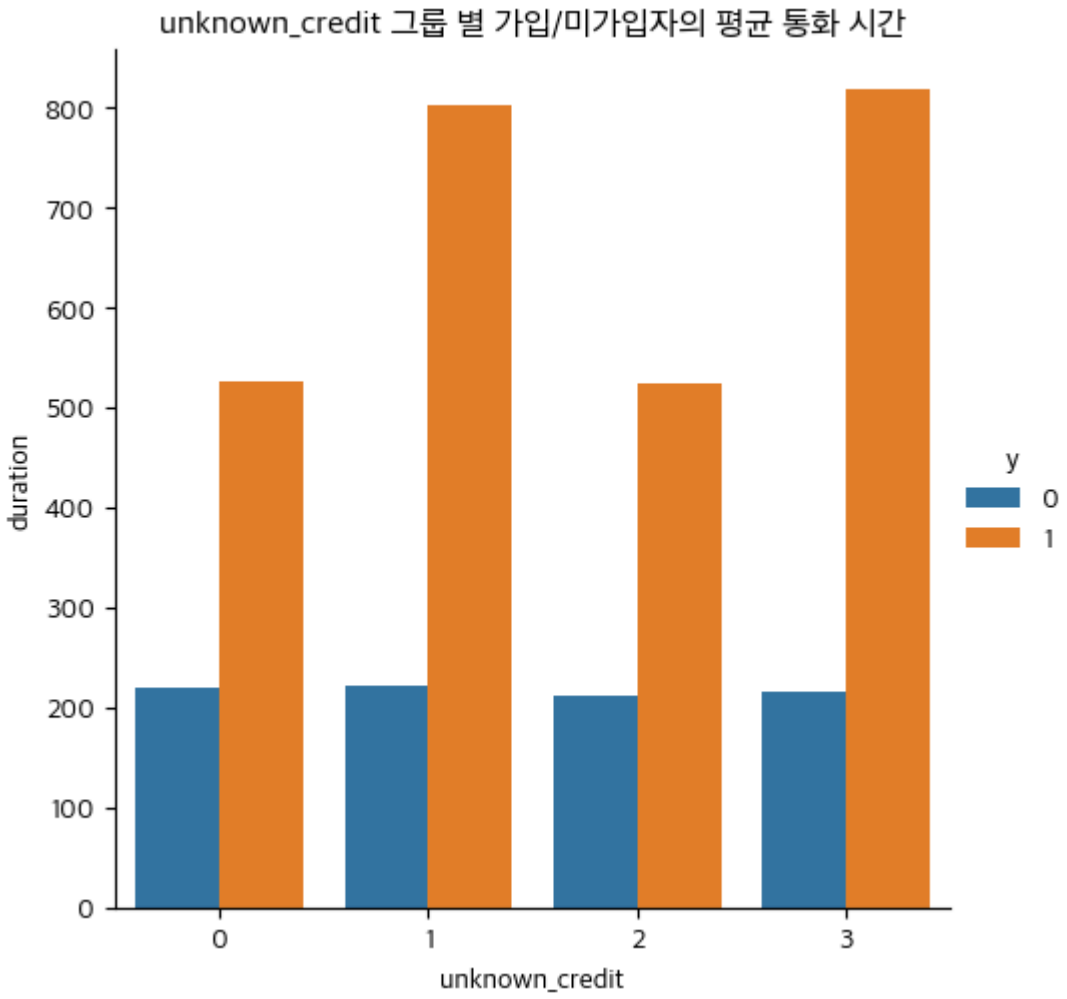
결론: Unknown의 개수는 가입률에 유의미한 영향을 미치지 않는다.

추가 가설: 통화 시간과 횟수가 Unknown 항목의 문제를 해결해줄 수 있다.

통화 시간(`duration`)

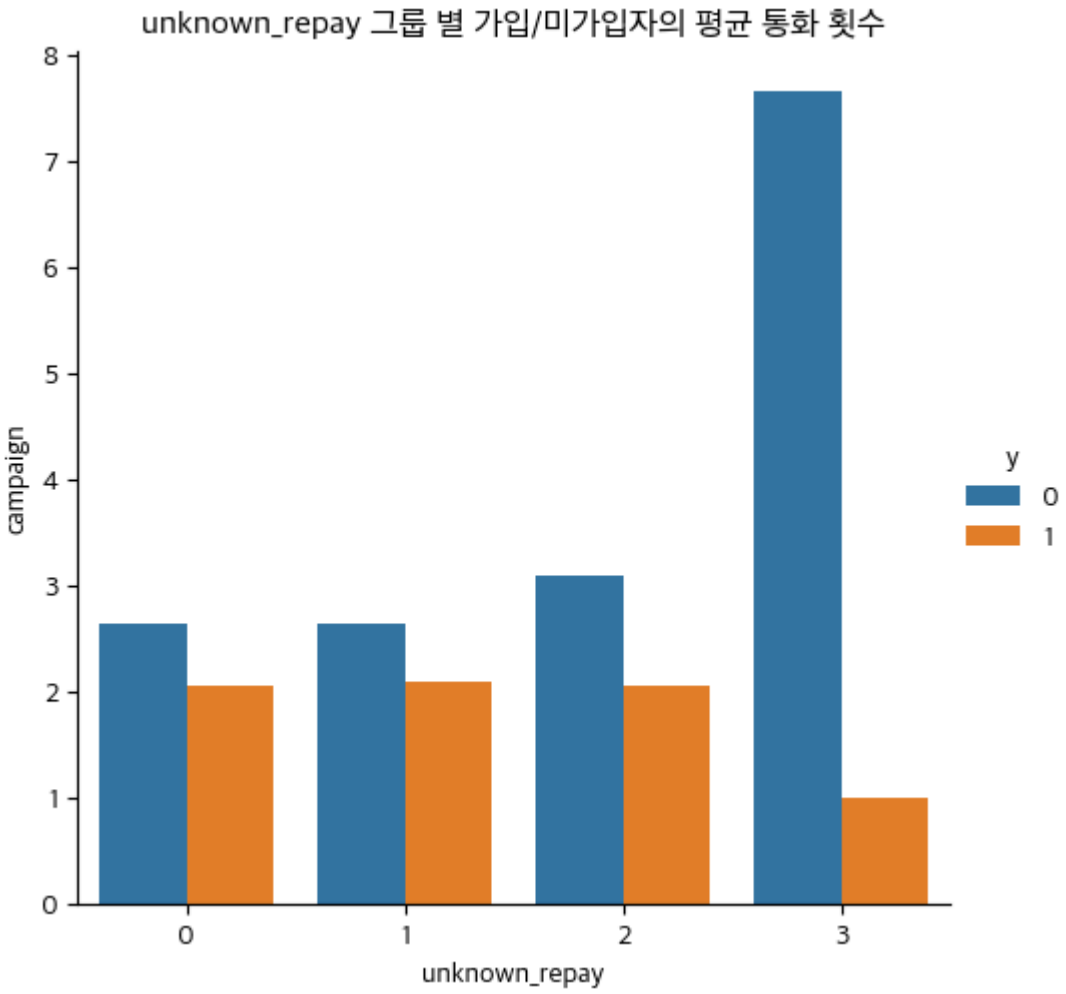

unknown_count 그룹 별 가입/미가입자의 평균 통화 시간
duration
700
600
500
400
300
200
100
0
0
1
2
3
4
5
unknown_count
y
0
1

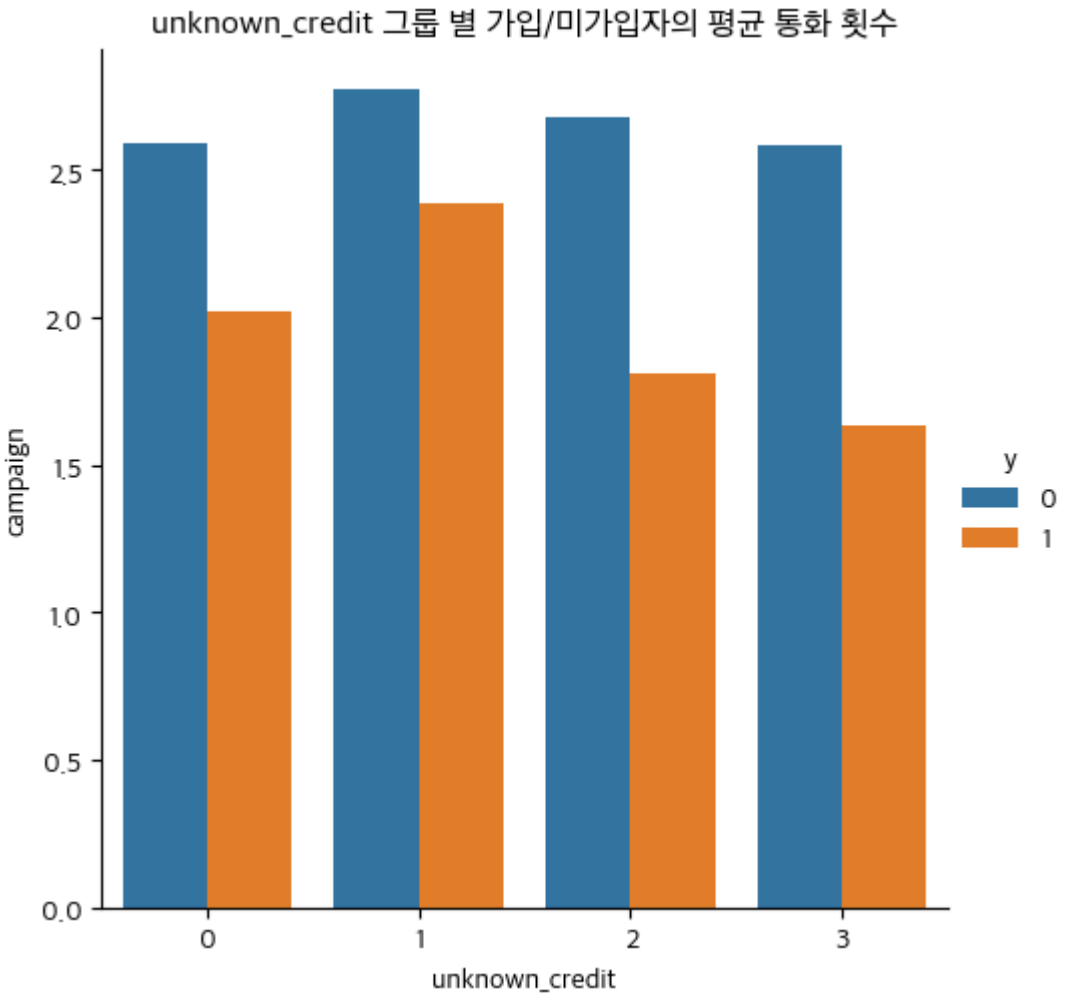

unknown_repay 그룹 별 가입/미가입자의 평균 통화 시간
duration
500
400
300
200
100
0
0
1
2
3
unknown_repay
y
0
1



- 그룹별 막대그래프와 산점도를 통해 Unknown 항목의 개수가 똑같아도, 통화 지속시간이 더 긴 그룹이 가입률이 높은 것으로 나타났다.

통화 횟수(`campaign`)

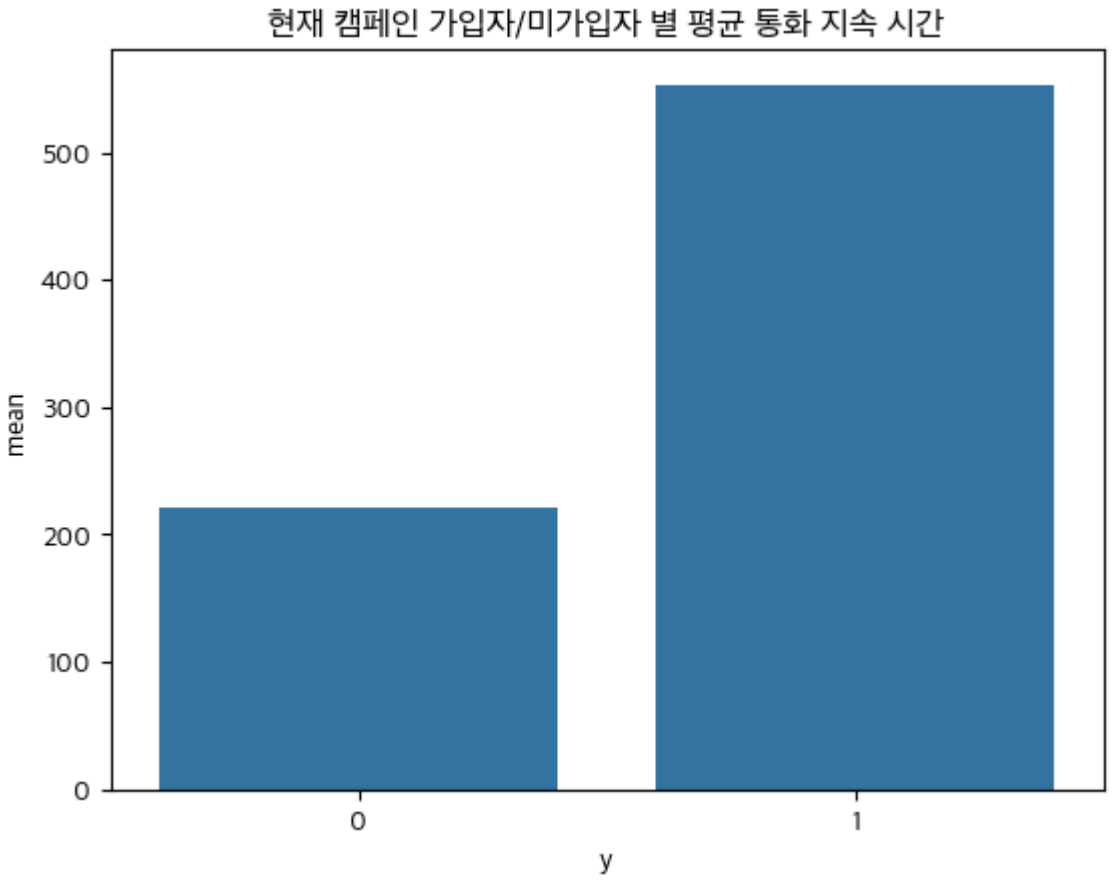

unknown_repay 그룹 별 가입/미가입자의 평균 통화 횟수
campaign
unknown_repay
y
0
1
0
1
2
3
4
5
6
7
8

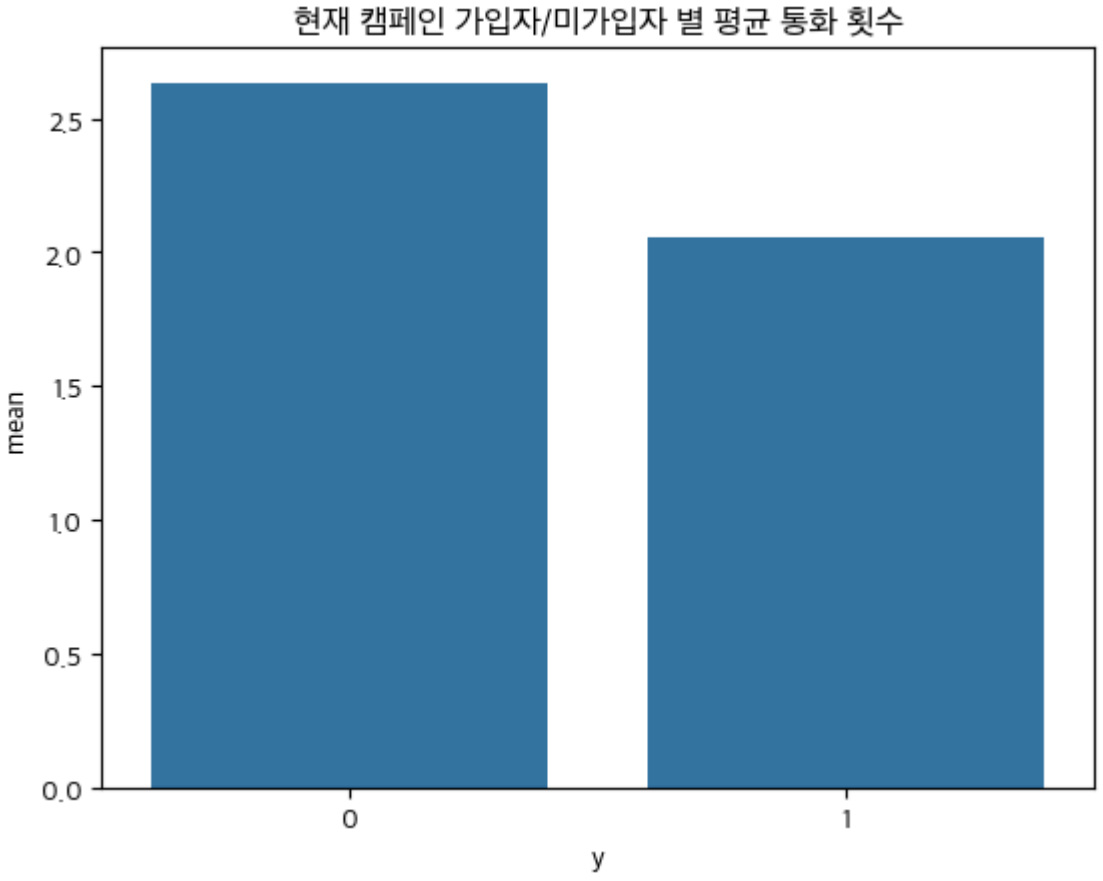

unknown_credit 그룹 별 가입/미가입자의 평균 통화 횟수
campaign
unknown_credit
y
0
1
0.0
0.5
1.0
1.5
2.0
2.5
0
1
2
3

가입까지의 연락 횟수는 가입자보다 미가입자가 더 많은 것으로 나타났다.

결론: 통화 시간은 긍정적인, 통화 횟수는 부정적인 양상을 보인다.

가설 확대: 통화 시간과 횟수의 양상은 **Unknown**에 국한되지 않고 전체적인 데이터에서도 관찰될 것이다.





결론: 통화 시간과 횟수가 가지는 가입률과의 상관성은 전체 데이터에서도 유사하게 드러난다.

이전 캠페인

- `pdays`, `previous`, `poutcome` 칼럼은 직전 캠페인과의 시간 간격(일 단위), 연락 횟수와 그 결과를 나타낸다.
- `poutcome`의 응답 중 nonexistent가 35,551명으로, 현재 캠페인에서 처음으로 연락한 잠재 고객이 대부분이다.

칼럼 보완

- `poutcome`의 failure 항목은 `pdays`가 999인 경우와 아닌 경우가 모두 존재하여, 해석과 학습에 혼동을 줄 수 있다.
- failure 중 `pdays`가 999인 경우와 아닌 경우를 `poutcome`에서 구분하도록 하였다.

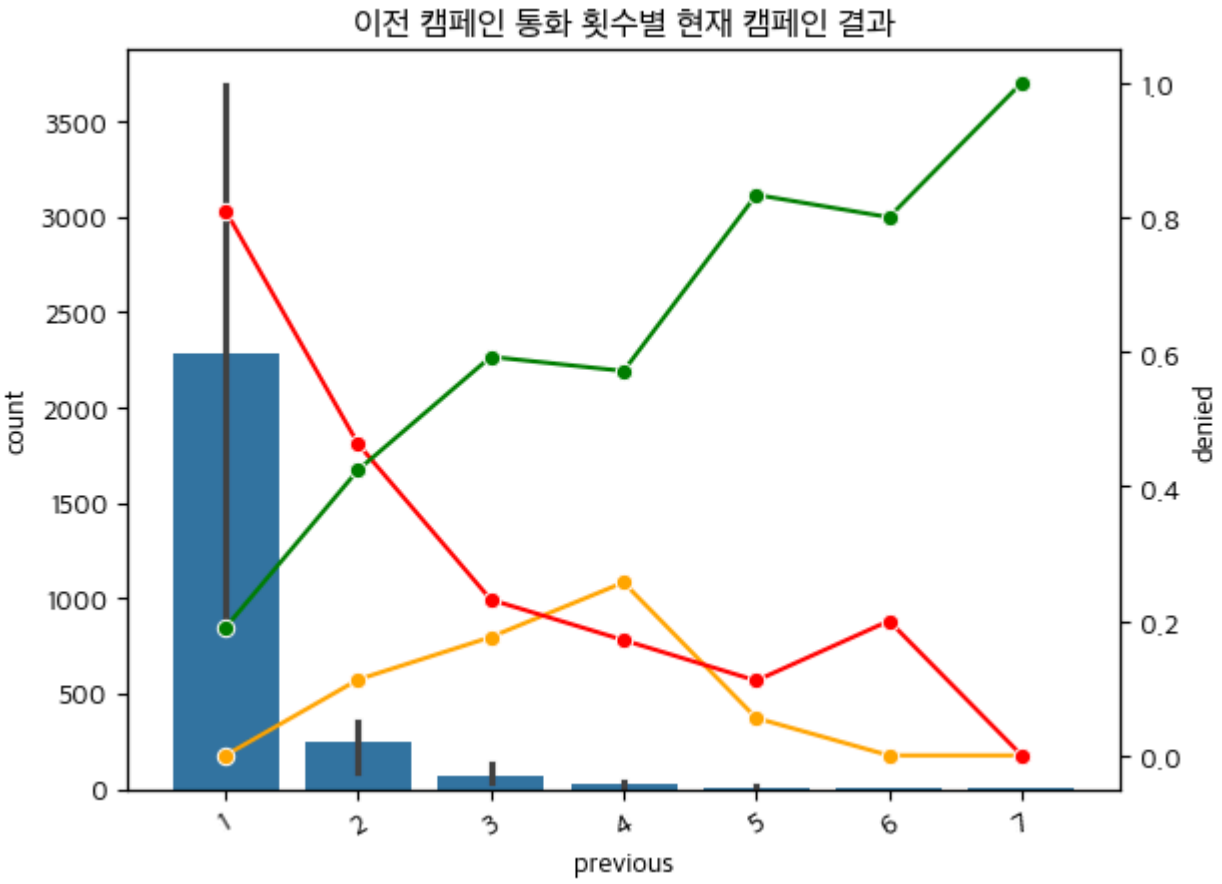
기존 칼럼 구성

이전 캠페인 이력	poutcome	previous	pdays
없음(첫 방문)	nonexistent	0	999
있음(예전에 시도 후 실패)	failure	!= 0	999: 부재중
있음(예전에 시도 후 실패)	failure	!= 0	!=999: 전화 후 거절
있음	success	!= 0	!= 999

수정 후 칼럼 구성

이전 캠페인 이력	poutcome	previous	pdays
없음(첫 방문)	nonexistent	0	999
있음(연락 실패)	missing	!= 0	999
있음(연락 성공 후 거절)	denied	!= 0	!=999
있음	success	!= 0	!= 999

가설 1: 이전 캠페인의 통화 데이터에서 도움이 되는 정보를 찾을 수 있다.



- 이전 캠페인의 통화 횟수

1. 통화 횟수 자체가 일회성인 경우가 많다는 점과 함께 미루어 봤을 때, 통화 횟수가 가입률에 영향을 미친다고 확정 짓기는 어렵다.
2. 통화 횟수가 많을수록 가입률도 높은 것으로 확인되므로, 이전 캠페인 데이터와 현재 캠페인 데이터의 통화 횟수 관련 결론은 상충된다.

- 이전 캠페인의 통화 시간
  현재 캠페인 데이터에서는 통화 시간이 예측에 일정 부분 도움이 되는 것으로 보이는데, 이전 캠페인 데이터에서 이를 검증할 수 없다.

결론: 이전 캠페인 통화데이터에서 1차 학습을 진행할 수 없다.

가설 2: 이전 캠페인의 결과로 현재 캠페인의 결과 예측에 도움이 될 것이다.

poutcome	mean	sum	count	
0	denied	0.514085	73	142
1	missing	0.129440	532	4110
2	nonexistent	0.088324	3140	35551
3	success	0.651129	894	1373

pdays	mean	sum	count	
0	1	0.638284	967	1515
1	999	0.092585	3672	39661

처음 만나는 잠재 고객보다 이전에 캠페인을 진행했던 그룹이 가입률이 높다.

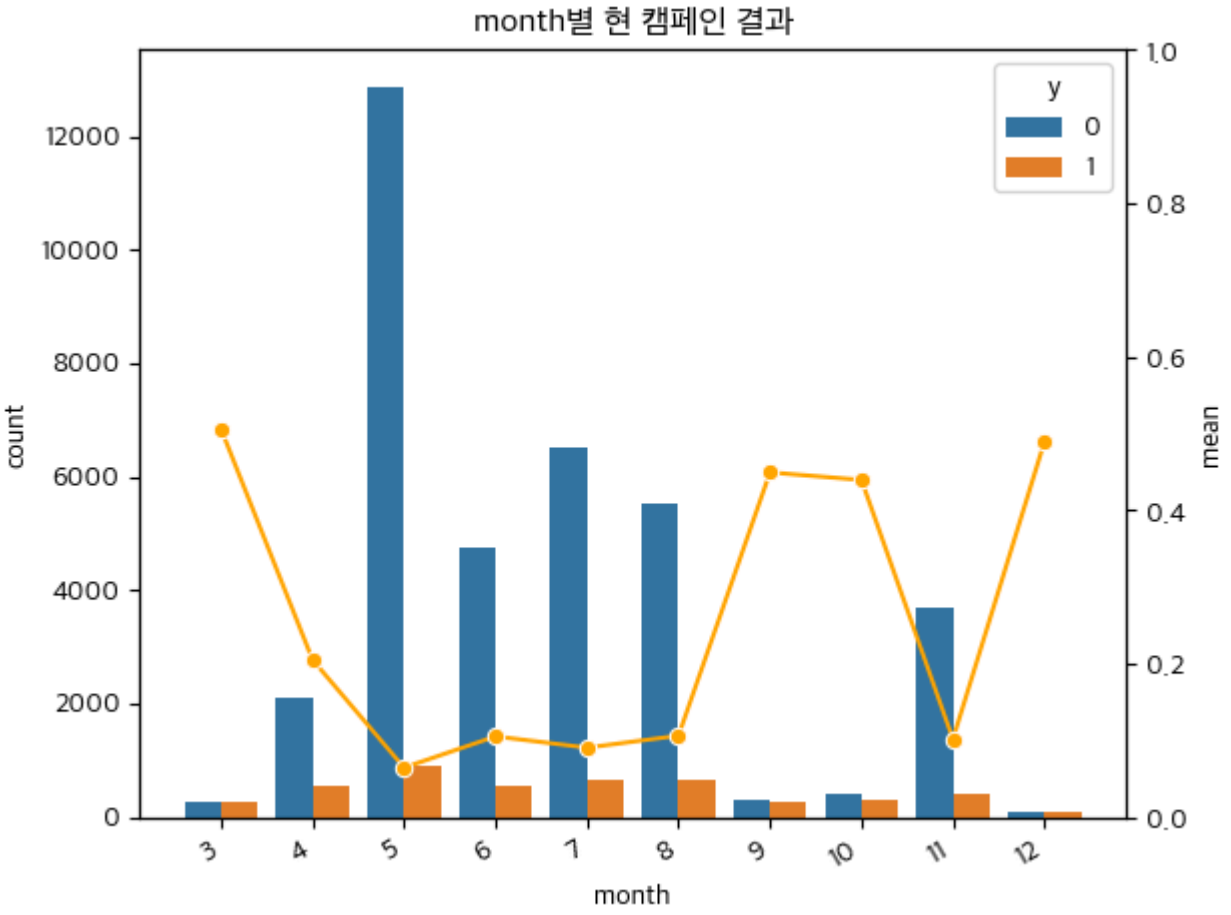
결론: 이전 캠페인의 결과가 현재 캠페인의 결과 예측에 도움이 될 수 있다.

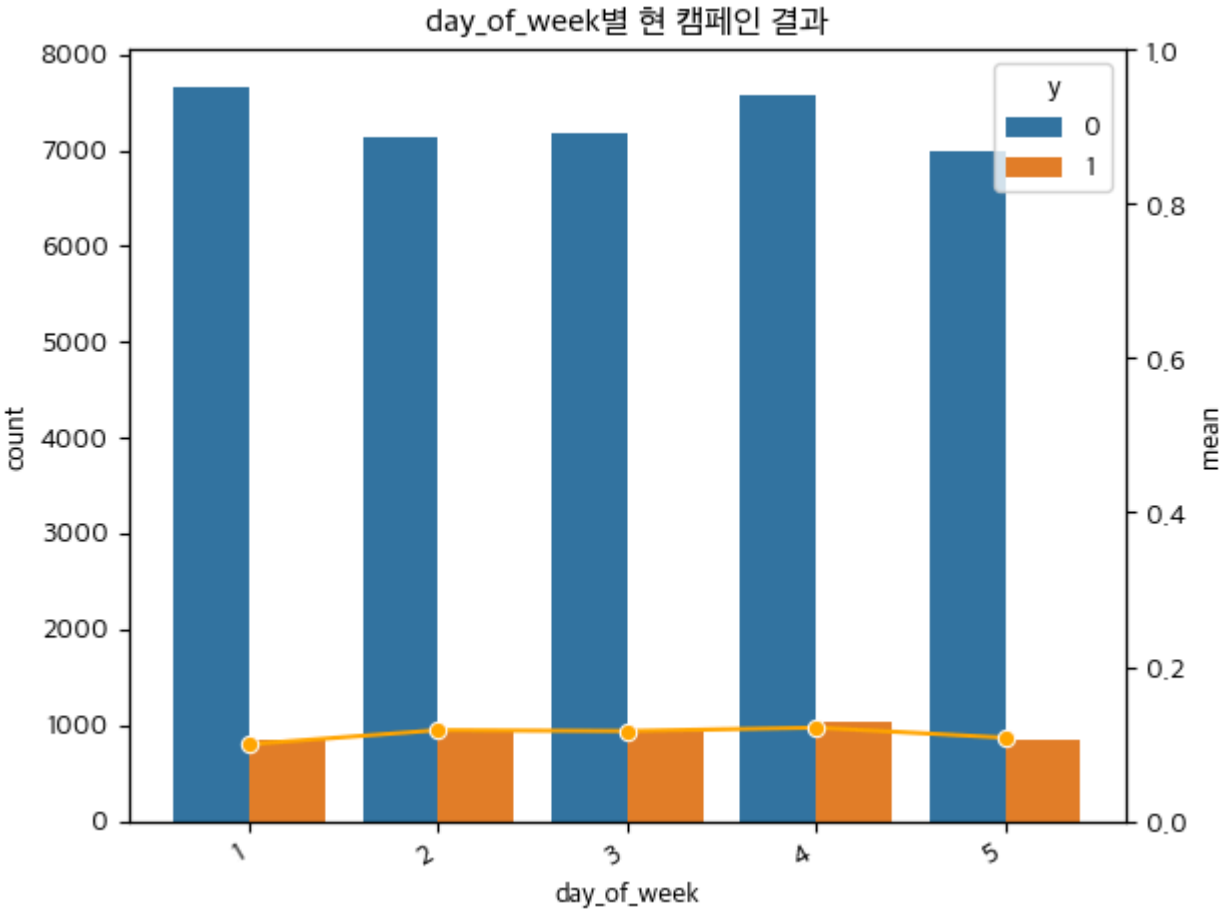
이전 캠페인 결론

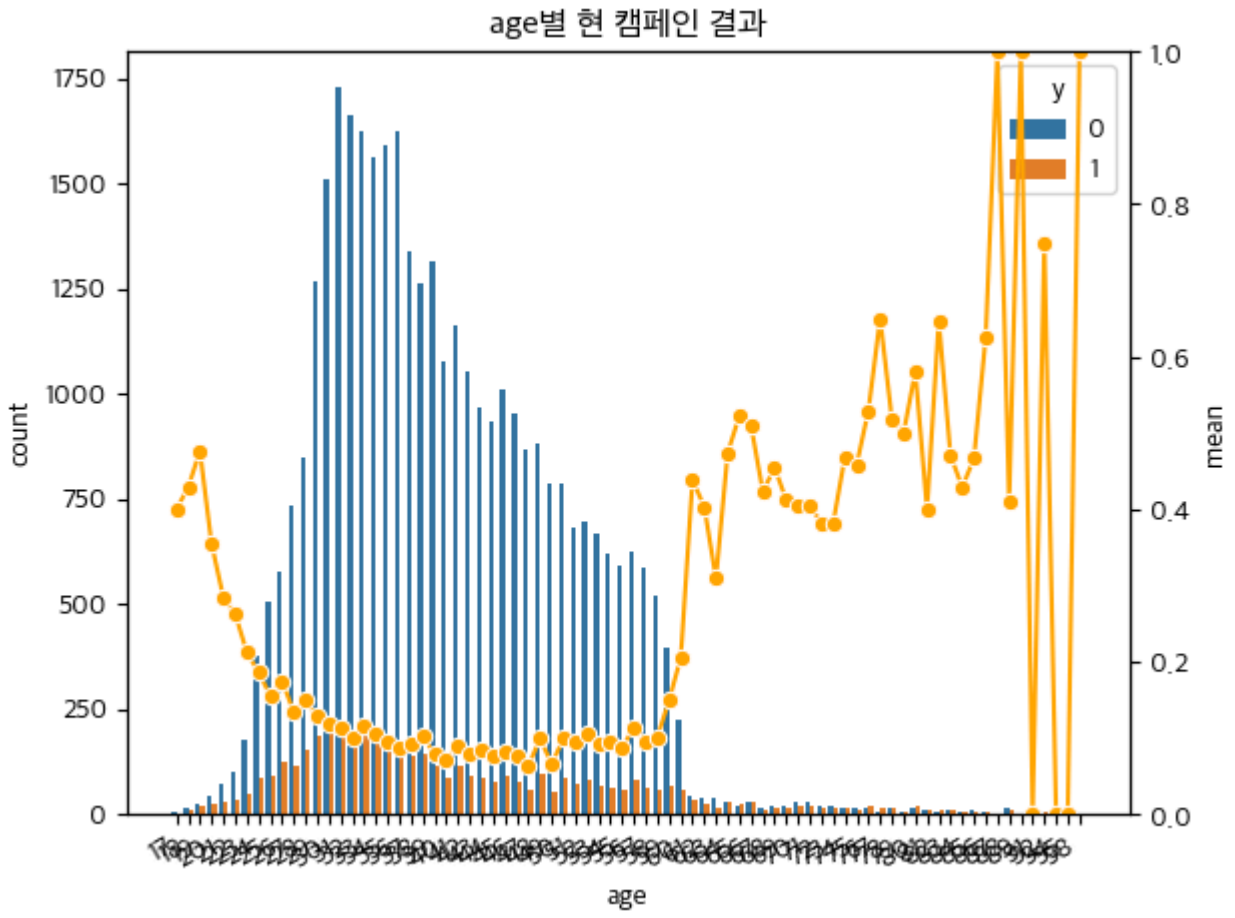
1. 통화 횟수는 결과 예측을 위한 일관성이 부족하다.
2. 통화 시간은 유의미한 긍정적인 상관관계를 가지는 것으로 보인다.
3. 이전 캠페인을 진행한 그룹이 그렇지 않은 그룹보다 가입률이 높다.

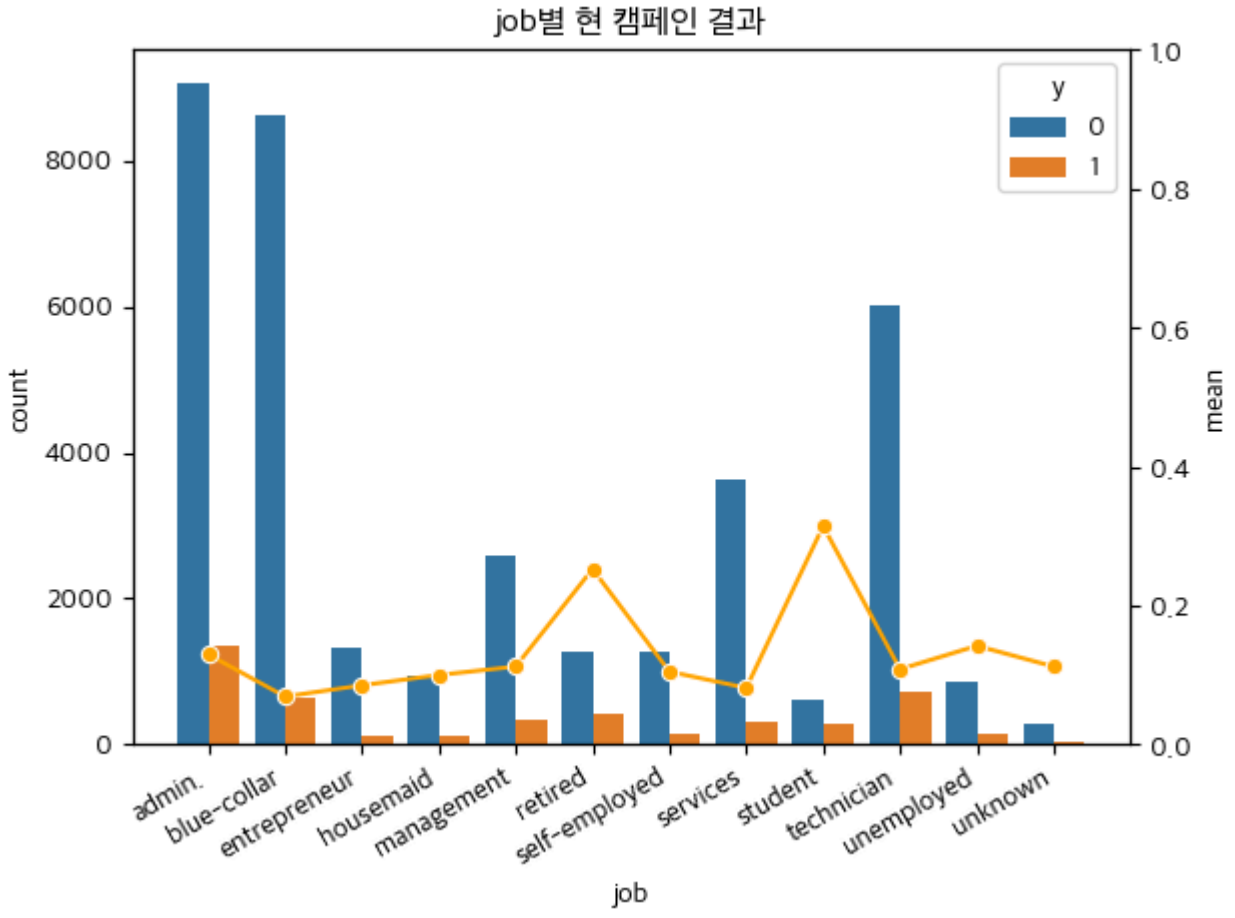
범주형 데이터 조사

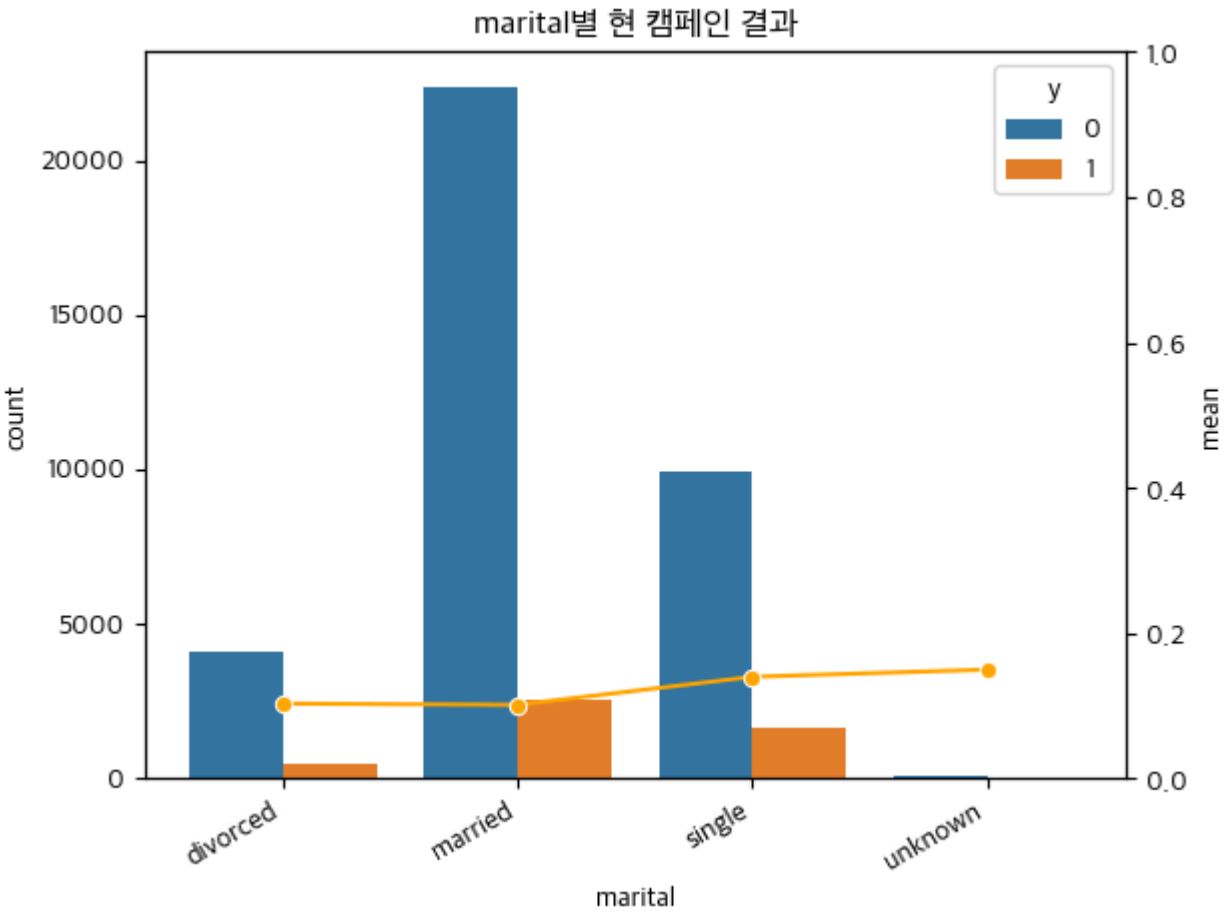
그룹별 현재 캠페인 결과

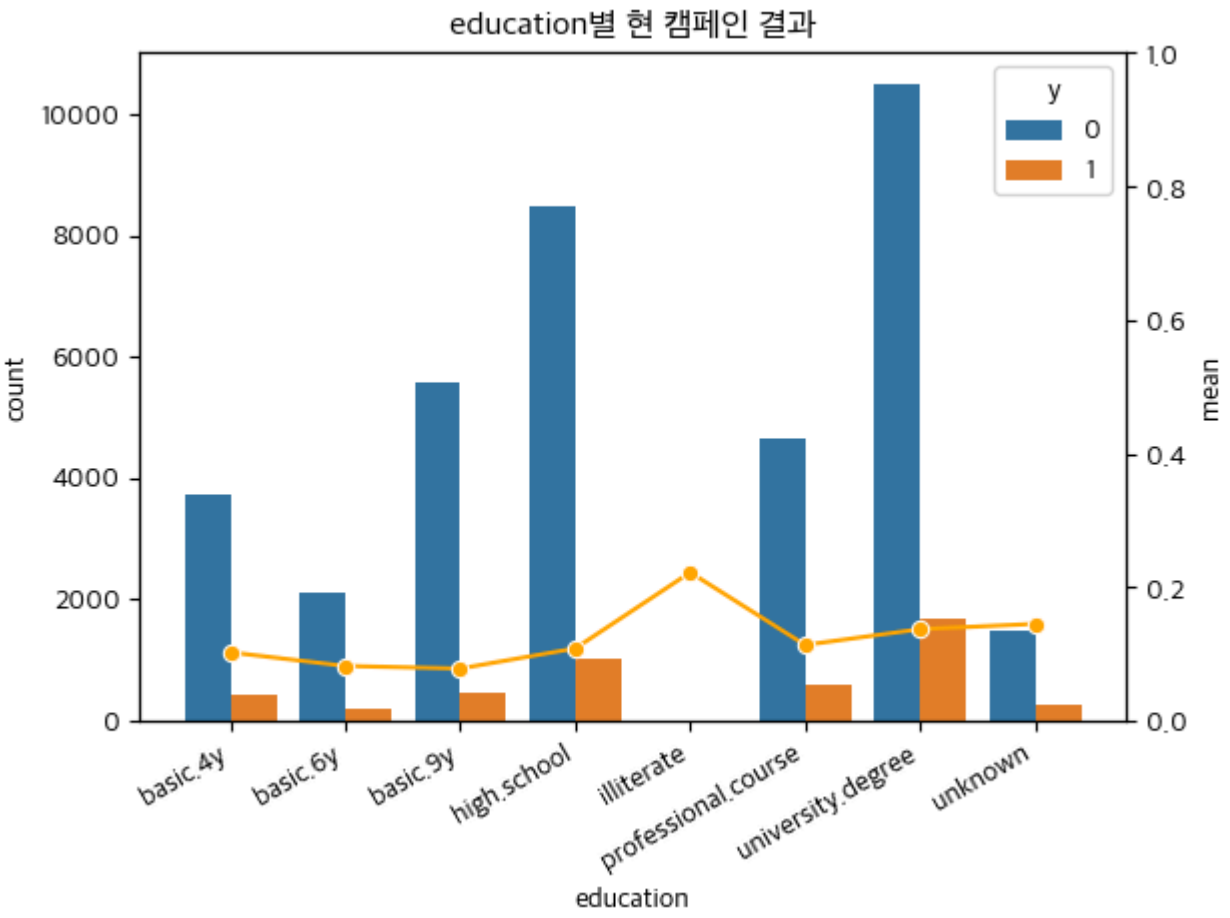

month별 현 캠페인 결과
y
0
1
count
mean
month
0
2000
4000
6000
8000
10000
12000
0.0
0.2
0.4
0.6
0.8
1.0
3
4
5
6
7
8
9
10
11
12

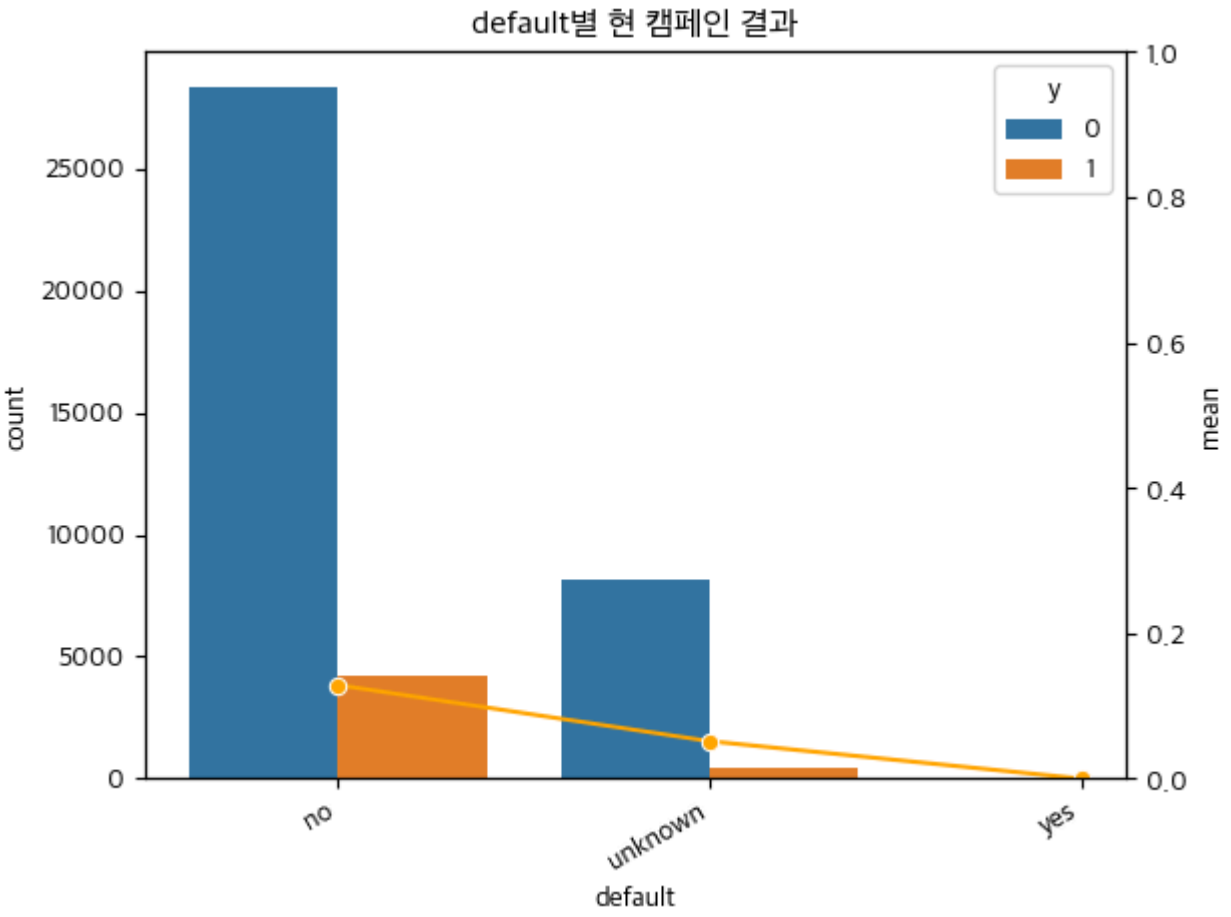

day_of_week별 현 캠페인 결과
y
0
1
count
mean
day_of_week
0
1000
2000
3000
4000
5000
6000
7000
8000
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5

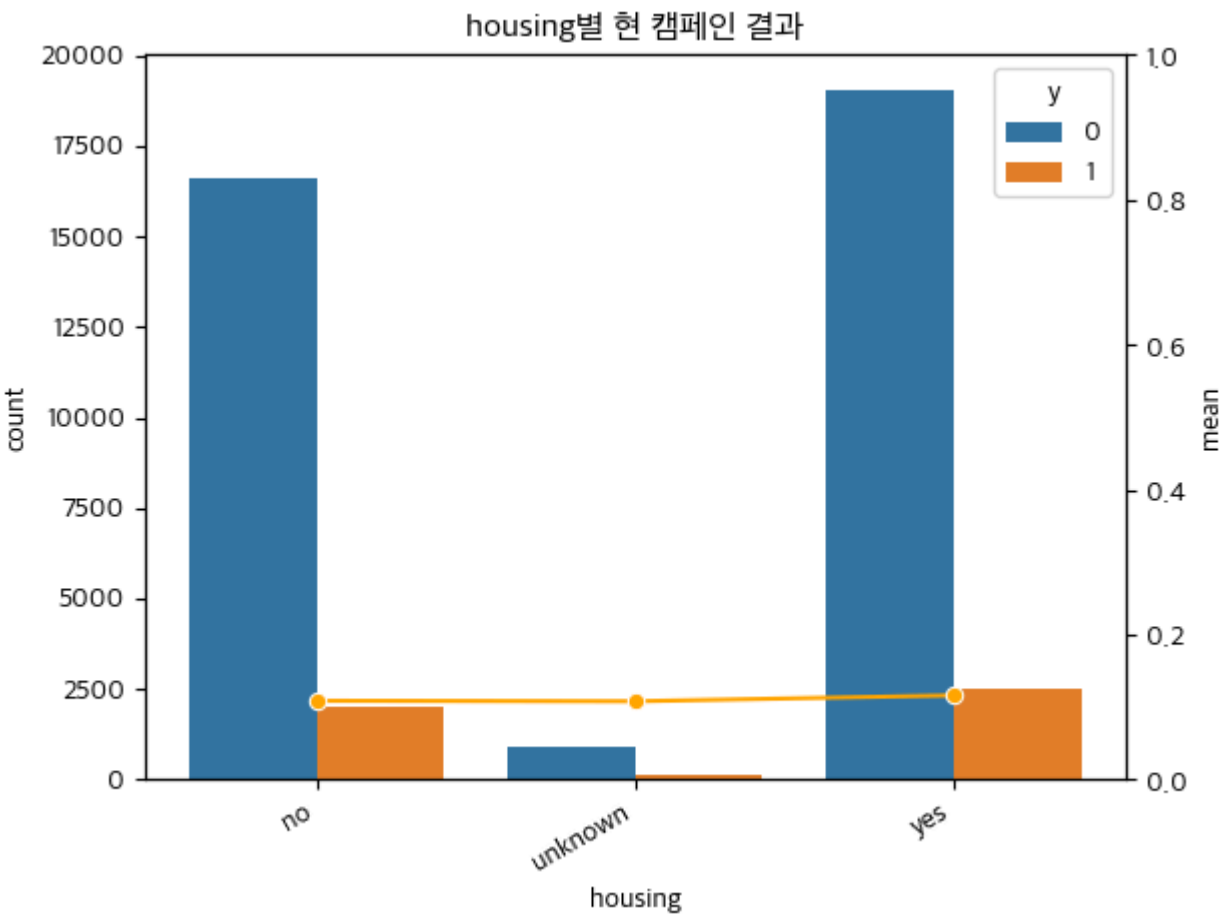

age별 현 캠페인 결과
count
mean
age
y
0
1
0
250
500
750
1000
1250
1500
1750
0.0
0.2
0.4
0.6
0.8
1.0

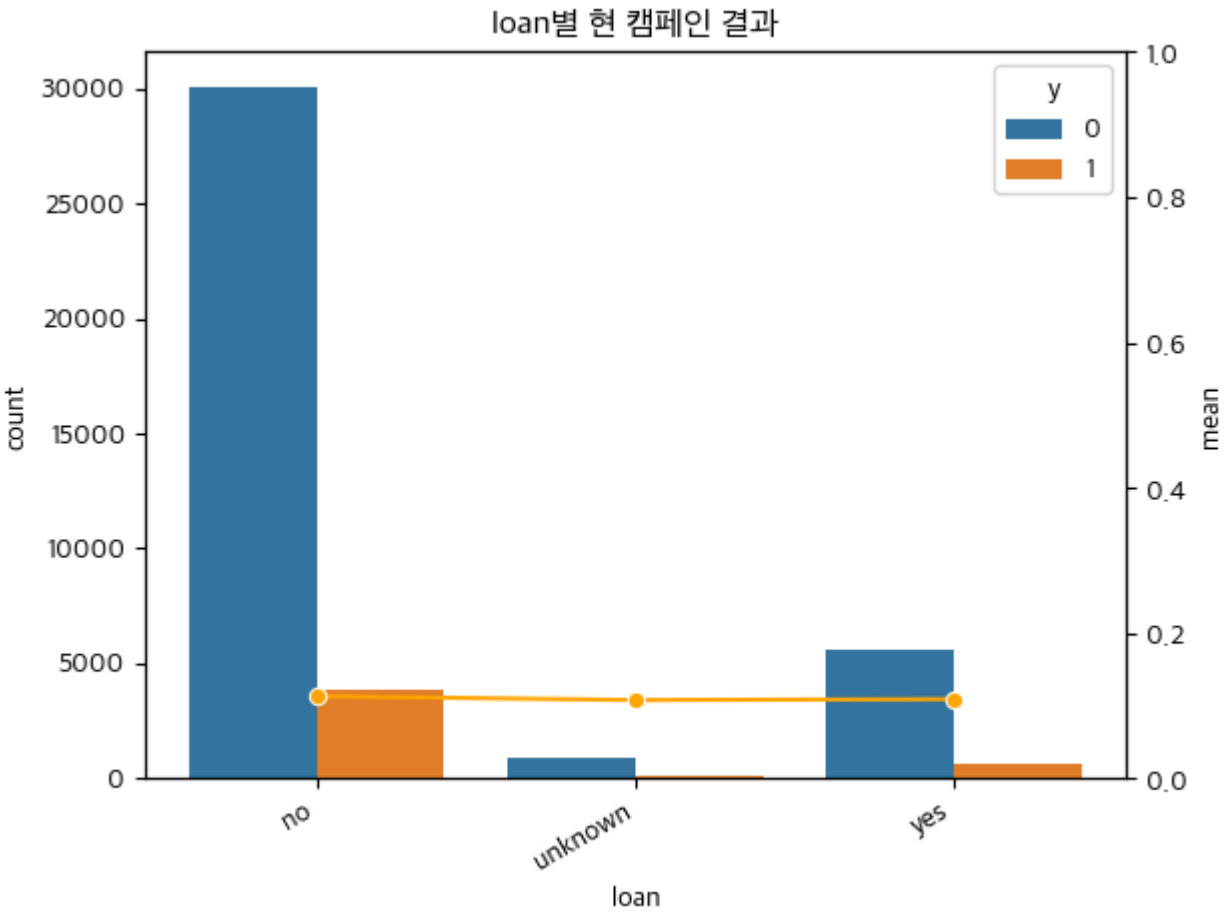

job별 현 캠페인 결과
count
mean
job
y
0
1
0
2000
4000
6000
8000
0.0
0.2
0.4
0.6
0.8
1.0
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown

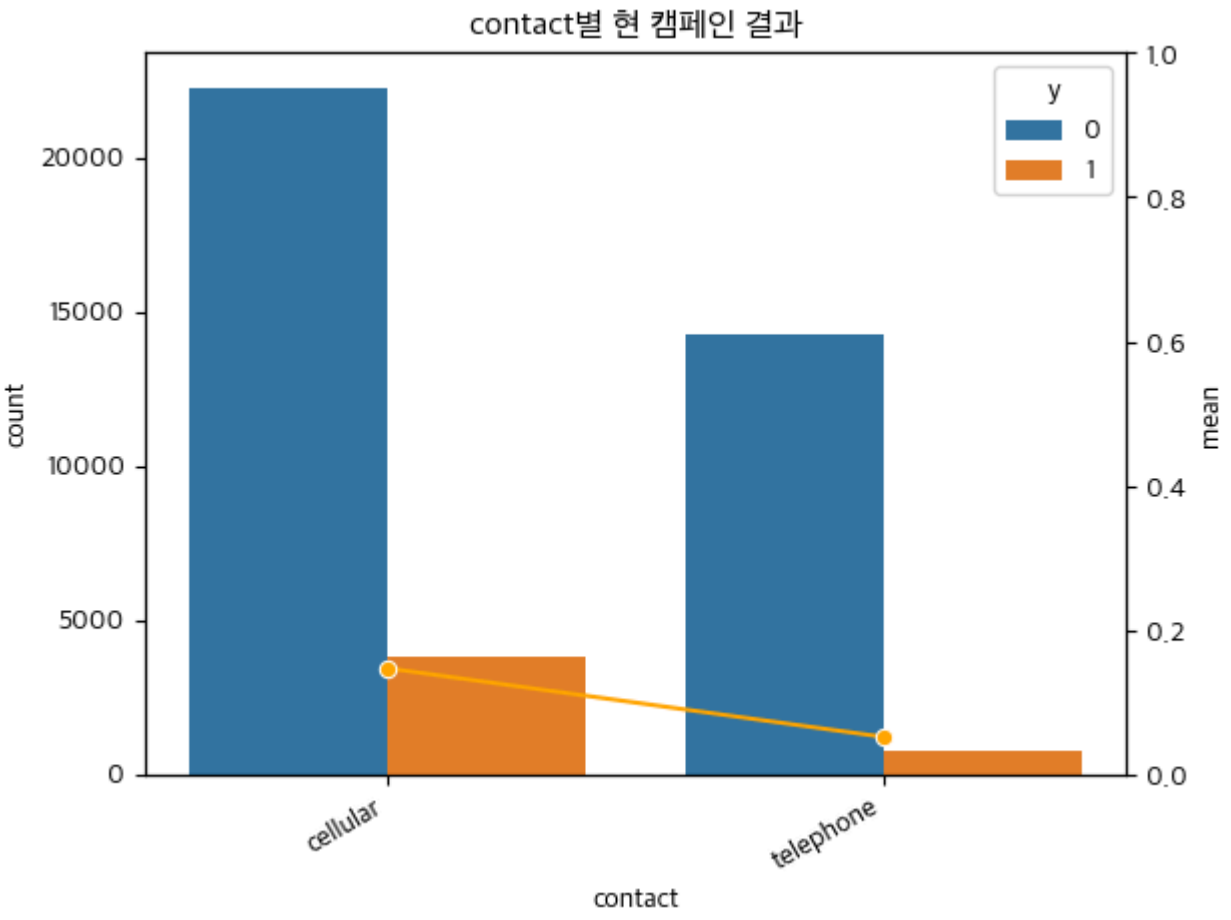

marital별 현 캠페인 결과
y
0
1
count
mean
20000
15000
10000
5000
0
1.0
0.8
0.6
0.4
0.2
0.0
divorced
married
single
unknown
marital


education별 현 캠페인 결과
y
0
1
count
mean
10000
8000
6000
4000
2000
0
1.0
0.8
0.6
0.4
0.2
0.0
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
education


default별 현 캠페인 결과
y
0
1
count
mean
25000
20000
15000
10000
5000
0
1.0
0.8
0.6
0.4
0.2
0.0
no
unknown
yes
default


housing별 현 캠페인 결과
y
0
1
count
mean
20000
17500
15000
12500
10000
7500
5000
2500
0
1.0
0.8
0.6
0.4
0.2
0.0
no
unknown
yes
housing





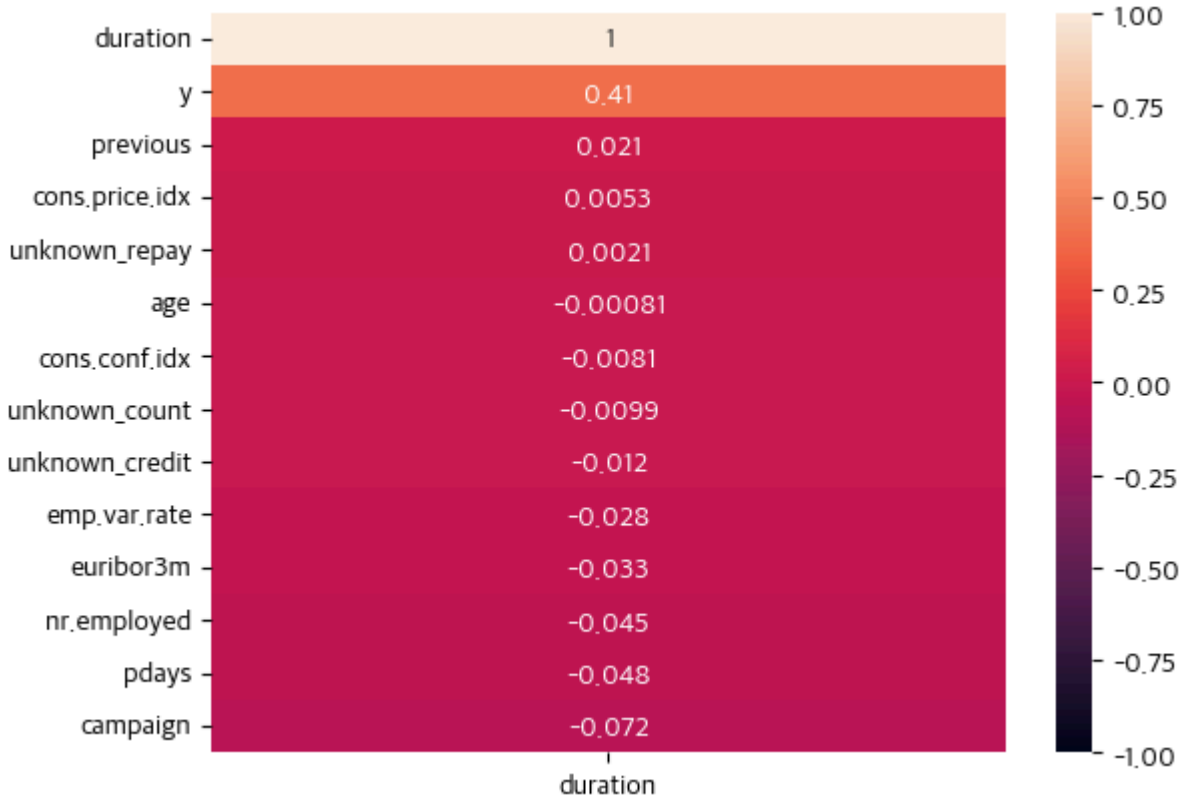
범주형 데이터 조사 결과

1. 통화 시간 외의 결정적인 지표를 식별하기 어려웠다.

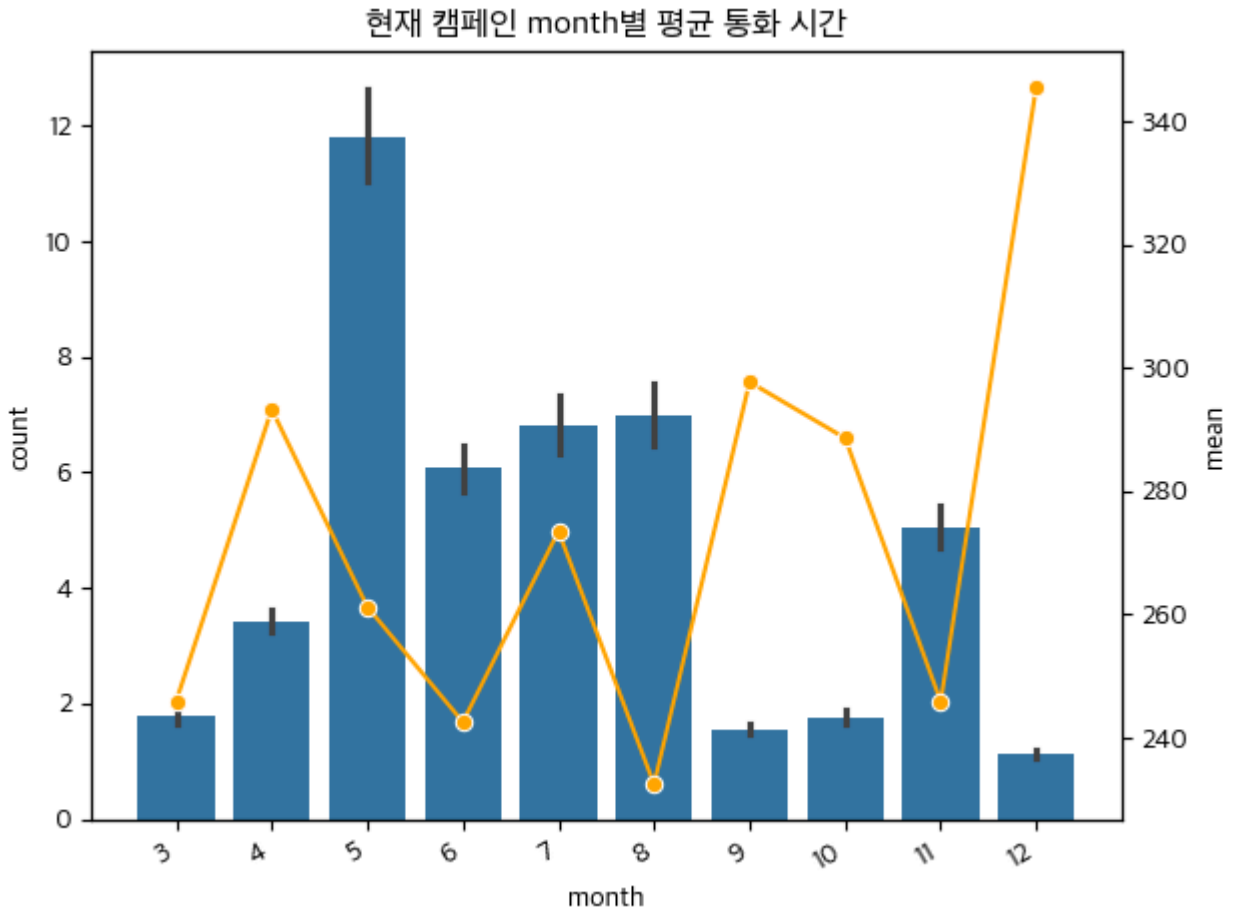
2. 통화 시간이 가입률을 예측할 수 있는 결정적인 지표로 해석된다.
3. 그러나 통화 시간은 캠페인 진행 후의 결과로, 통화 진행 전에는 통화 지속 시간을 알 수 없다는 단점이 있다.
4. 통화 지속 시간에 결정적인 영향을 미치는 요소 탐색이 추가적으로 필요하다.

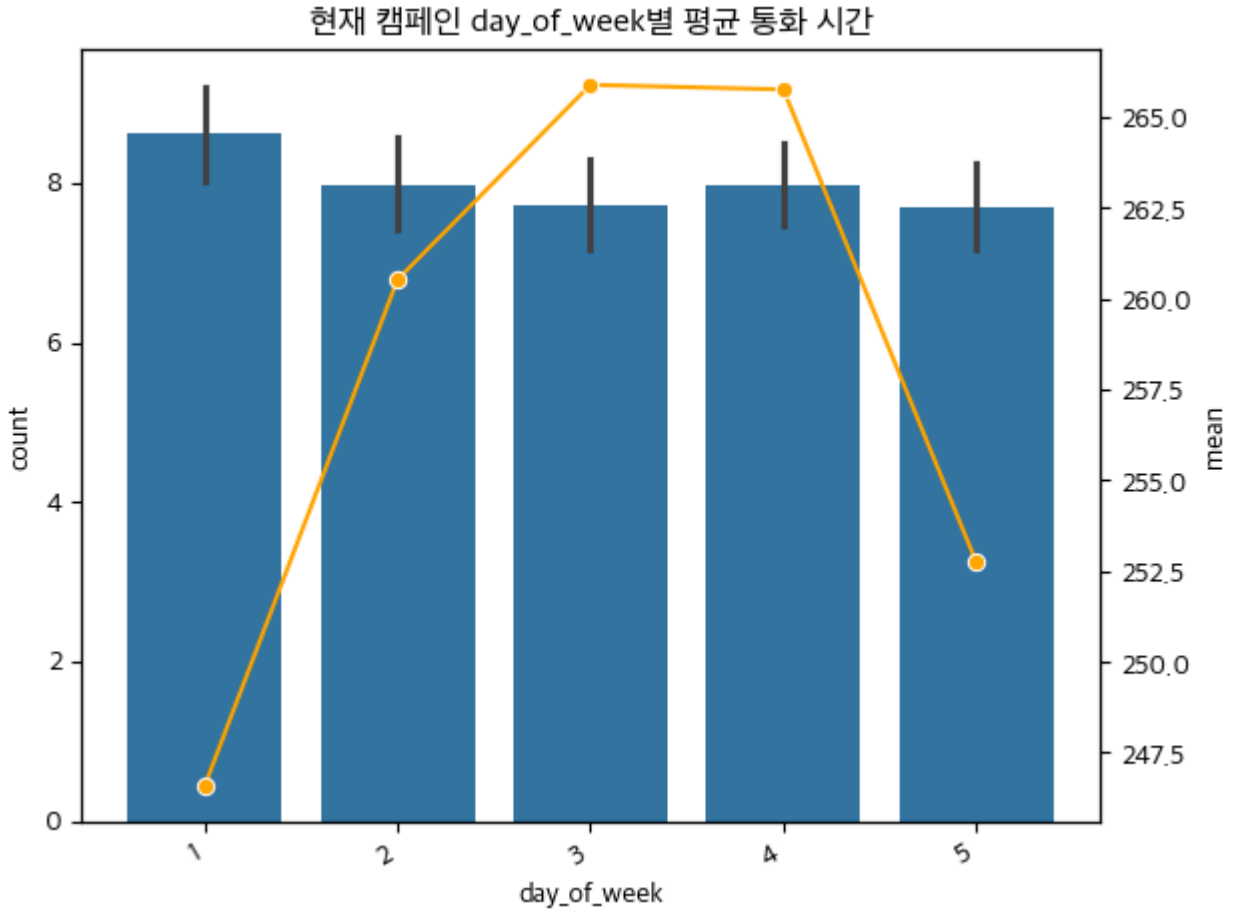
통화 시간 관련 지표 탐색

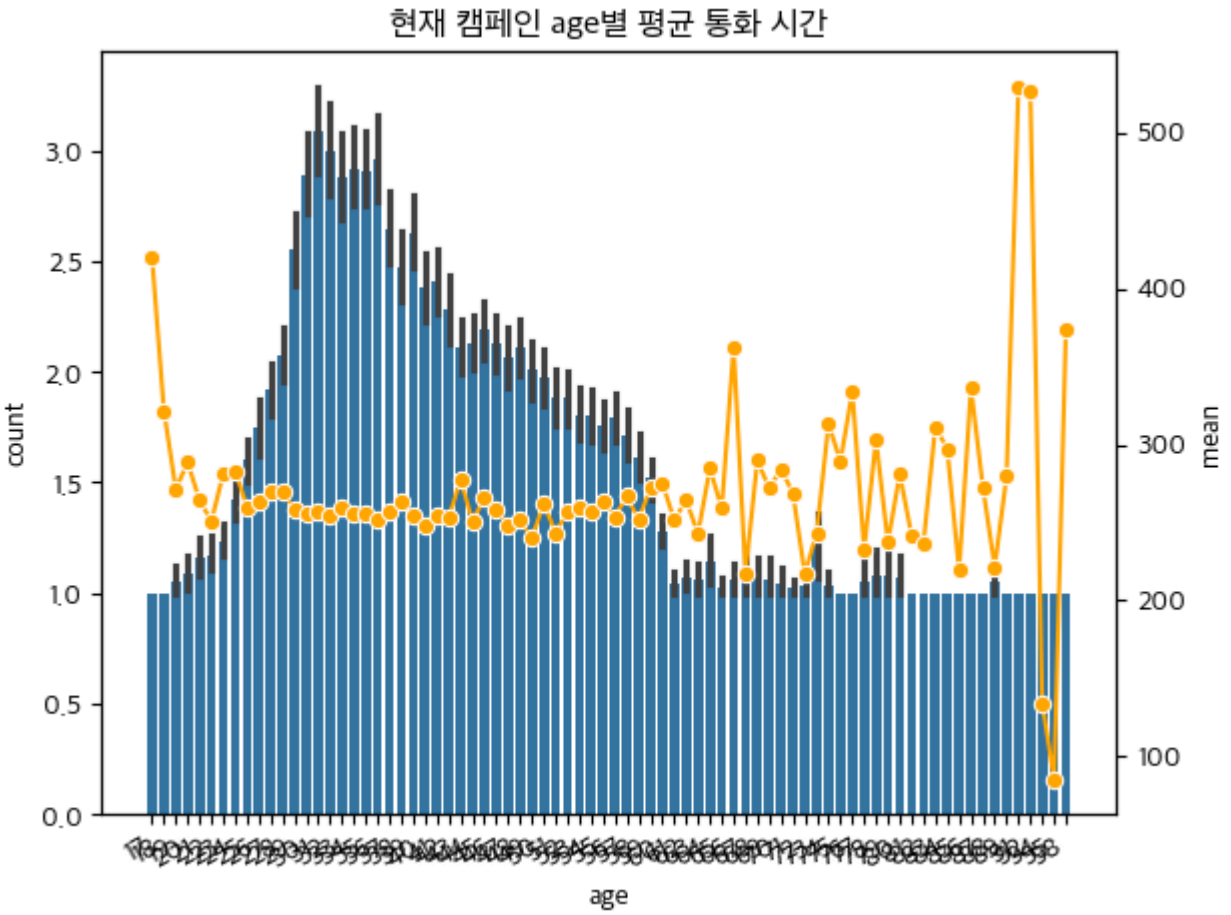
상관관계

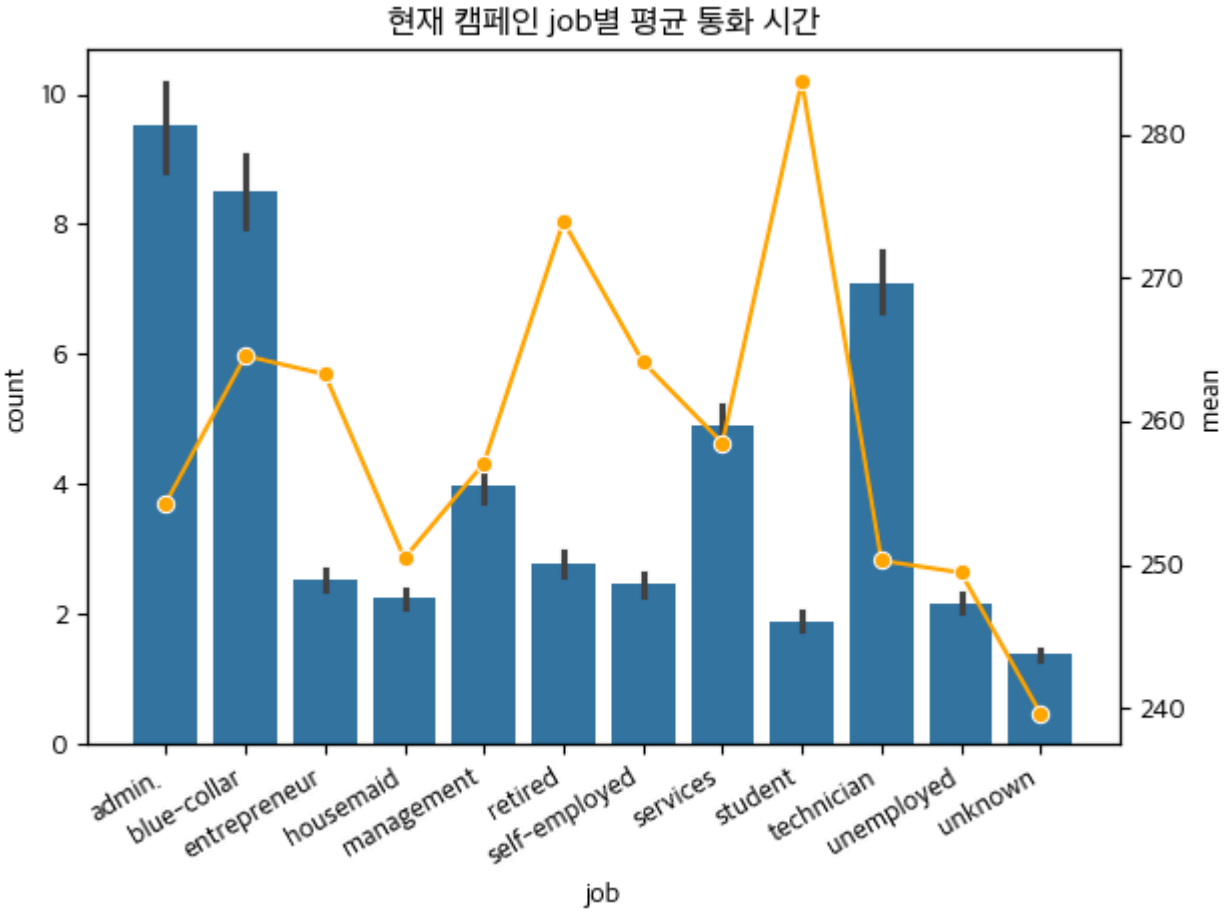


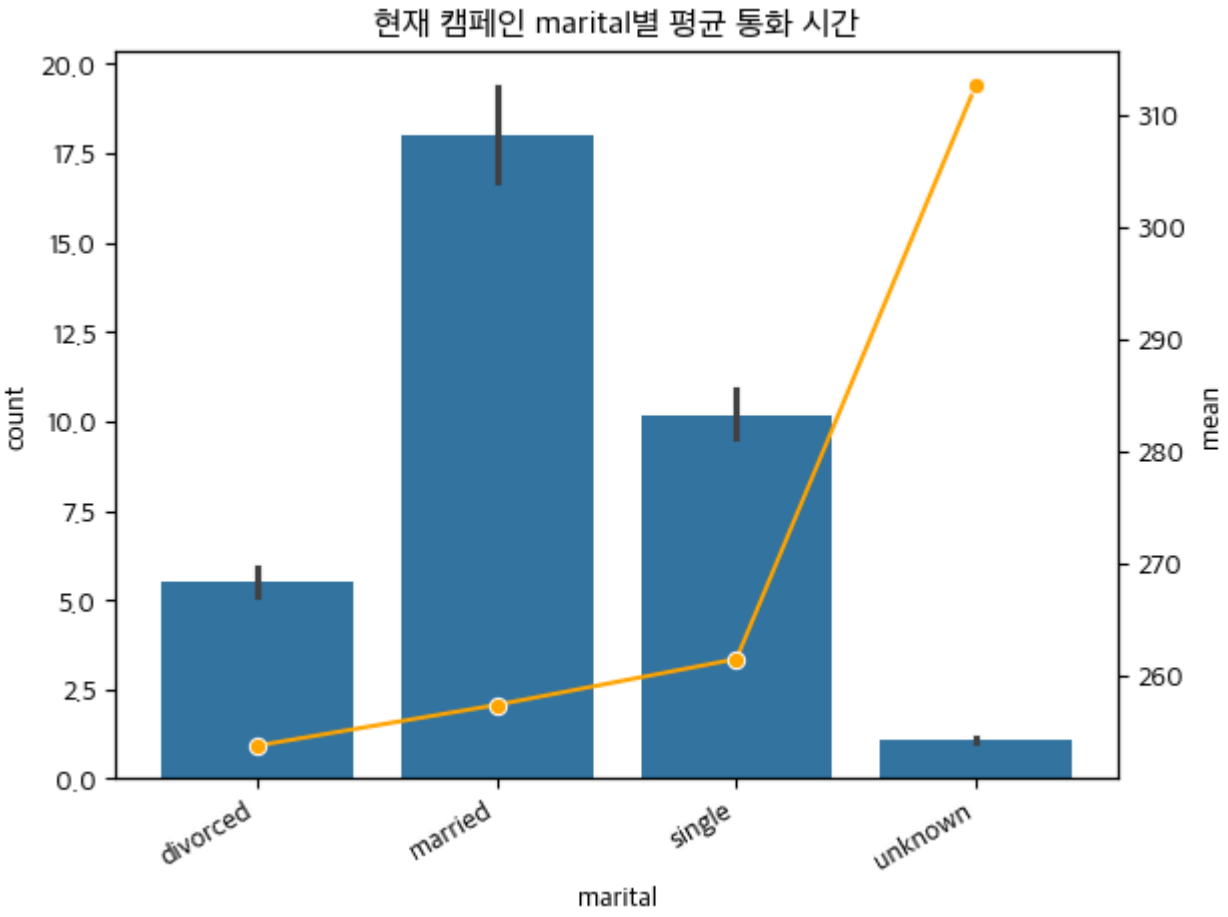
범주형 데이터

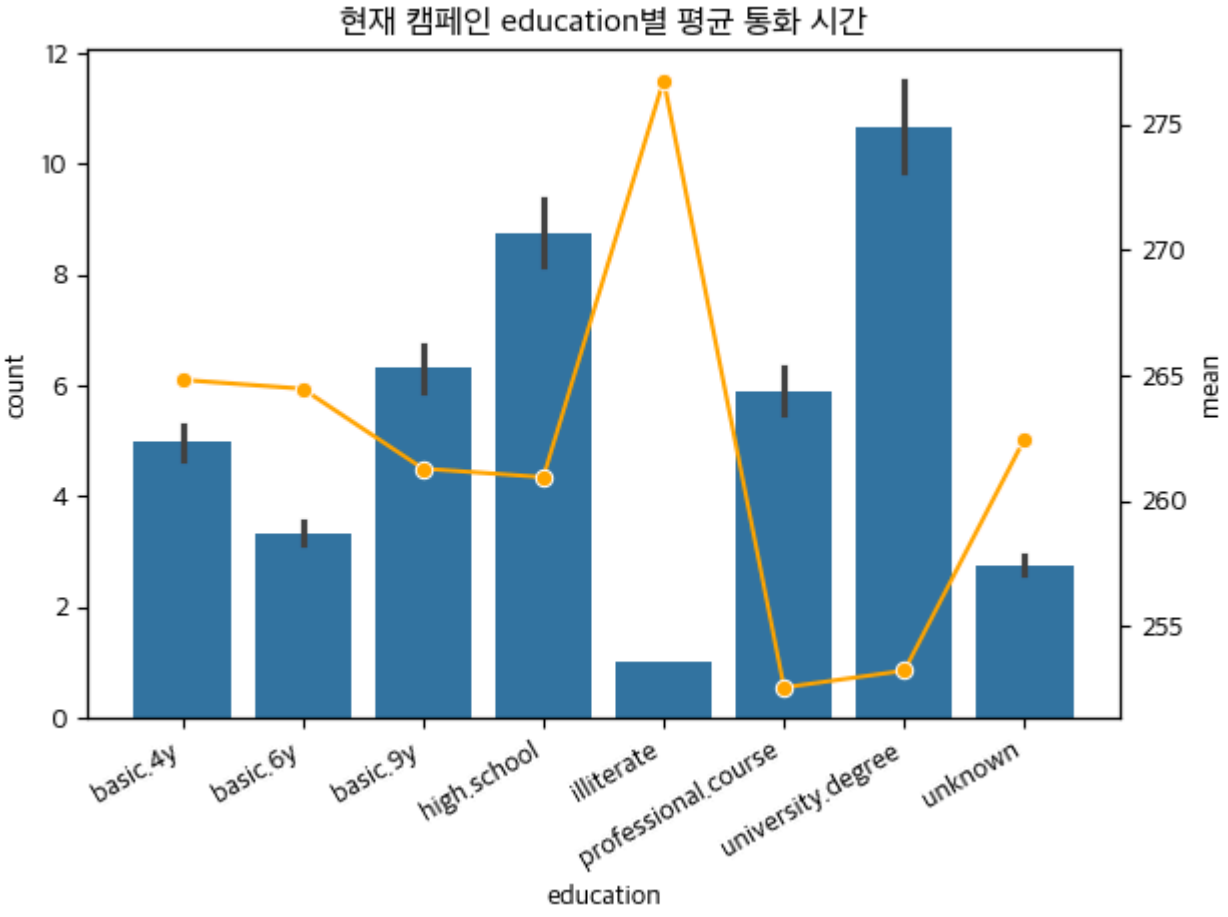

현재 캠페인 month별 평균 통화 시간
count
mean
month
0
2
4
6
8
10
12
240
260
280
300
320
340
3
4
5
6
7
8
9
10
11
12

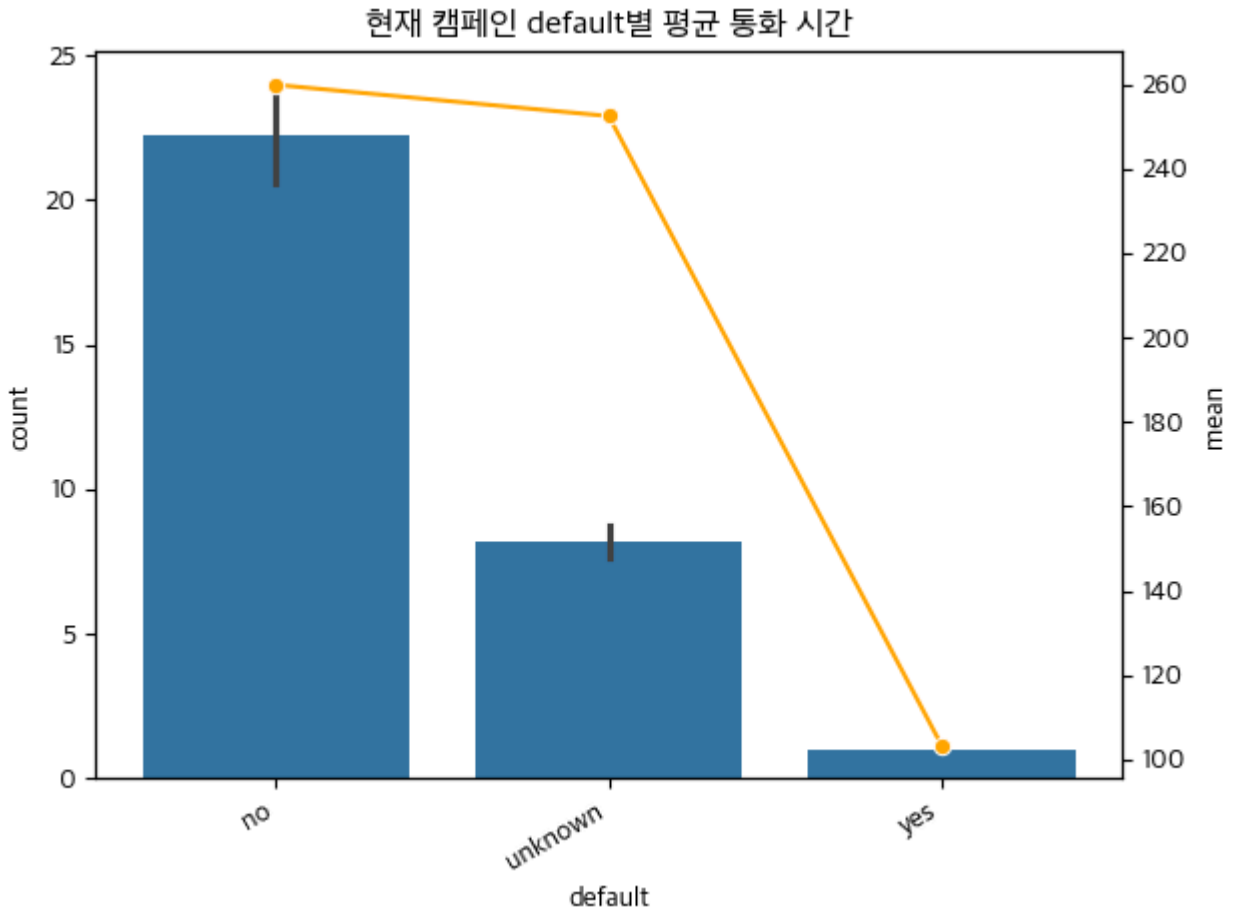

현재 캠페인 day_of_week별 평균 통화 시간
count
mean
day_of_week
0
2
4
6
8
247.5
250.0
252.5
255.0
257.5
260.0
262.5
265.0
1
2
3
4
5

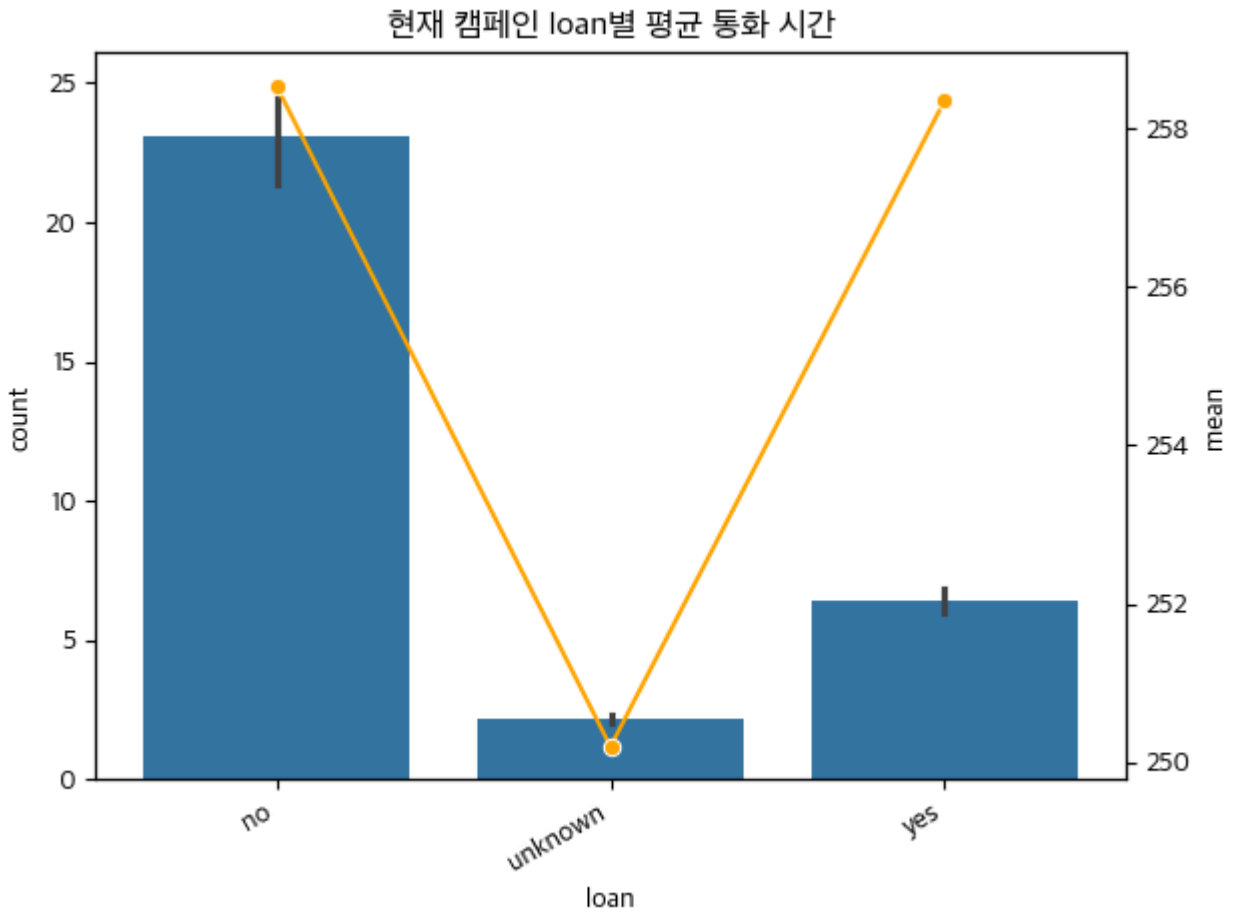

현재 캠페인 age별 평균 통화 시간
count
mean
age
0.0
0.5
1.0
1.5
2.0
2.5
3.0
100
200
300
400
500

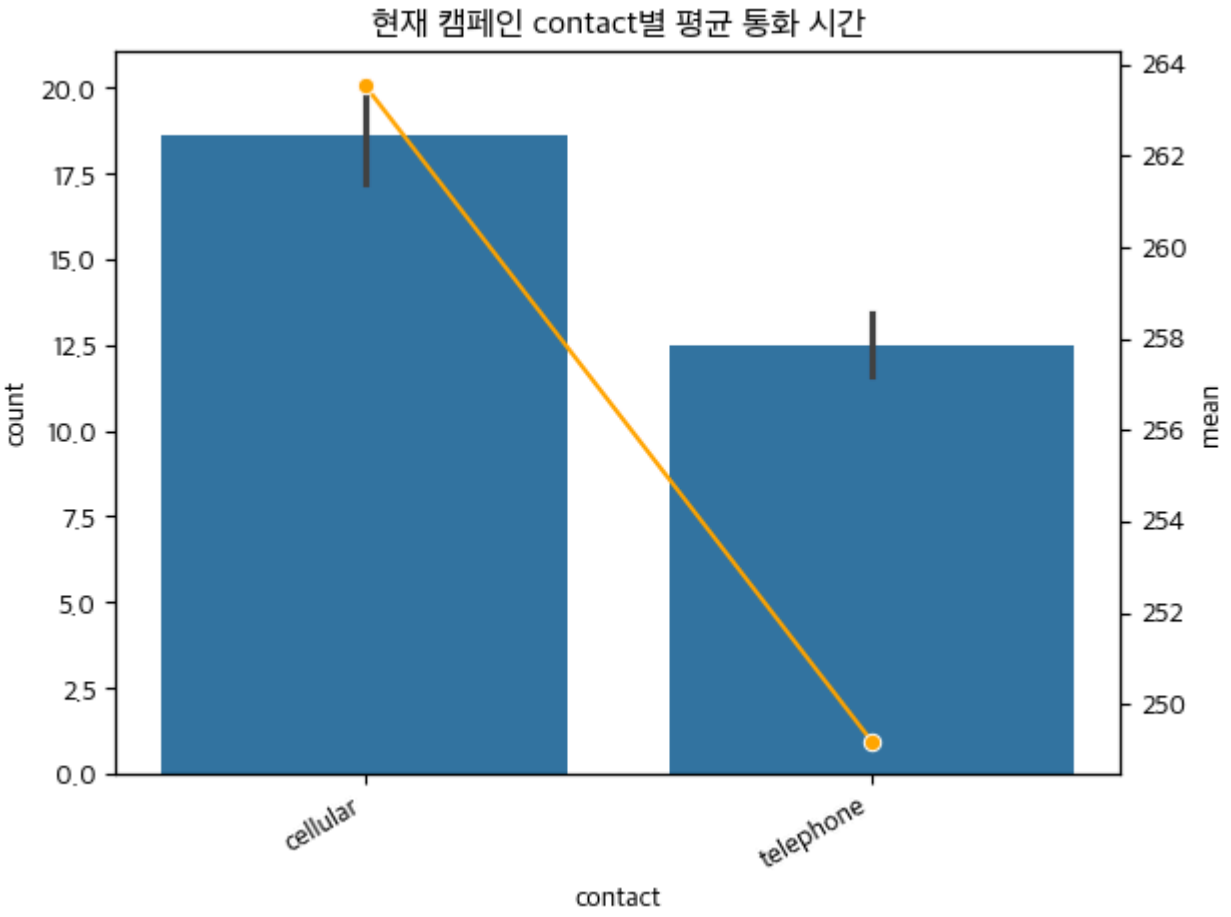

현재 캠페인 job별 평균 통화 시간
count
mean
job
0
2
4
6
8
10
240
250
260
270
280
admin.
blue-collar
entrepreneur
housemaid
management
retired
self-employed
services
student
technician
unemployed
unknown


현재 캠페인 marital별 평균 통화 시간
count
mean
marital
divorced
married
single
unknown
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
260
270
280
290
300
310


현재 캠페인 education별 평균 통화 시간
count
mean
education
basic.4y
basic.6y
basic.9y
high.school
illiterate
professional.course
university.degree
unknown
0
2
4
6
8
10
12
255
260
265
270
275


현재 캠페인 default별 평균 통화 시간
count
mean
default
no
unknown
yes
0
5
10
15
20
25
100
120
140
160
180
200
220
240
260


현재 캠페인 loan별 평균 통화 시간
count
mean
loan
no
unknown
yes
0
5
10
15
20
25
250
252
254
256
258



범주형 데이터 탐색 결과

1. 통화 시간이 긴 그룹은 대체적으로 데이터 양이 적은 편으로, 신뢰하기 어려운 정보이다.
2. 통화 시간 관련 파생 변수 없이 데이터 학습을 진행해야 한다.

---

2. 데이터 전처리

1. 이전 캠페인의 가설 2에서 얻은 결론을 토대로, 이전 캠페인을 진행했던 그룹 여부를 나타내는 칼럼을 생성한다.
2. 기타 범주형은 원핫인코딩으로 진행한다.

---

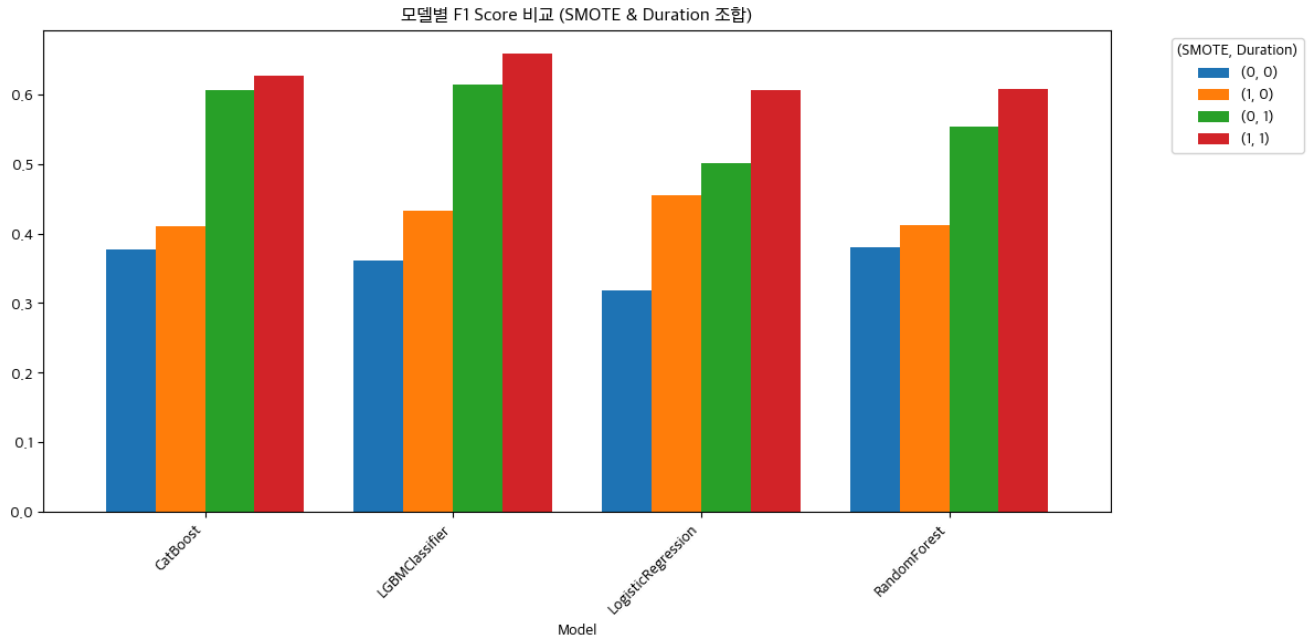
3. 실험 설계

1. 통화 시간을 제외하여 학습한 모델의 성능과 포함하여 학습한 모델의 성능을 비교하여, 통화 시간의 영향력을 파악한다. 실험 결과에 따라 수집하는 데이터를 개선하는 방안을 모색할 수 있다.
2. 데이터의 클래스 불균형을 보완하기 위해 SMOTE 적용 여부를 기준으로 각각의 실험 결과를 평가한다. 또한 StratifiedKfold를 진행한다.
3. y의 0·1을 분류하는 과제이므로, 학습 모델은 LogisticRegression, RandomForest, CatBoost, LGBMClassifier로 진행한다.

4. 성능 평가는 F1_score로 설정한다.

---

4. 실험 결과



결론

1. duration이 학습 데이터에 포함되었을 때 모델의 예상 성능이 향상되는 것을 확인할 수 있다.
2. 이를 통해, 지난 데이터의 통화 시간도 함께 데이터로 기록하여 미래 캠페인의 예측에 도움이 되어야 한다는 결론을 얻을 수 있다.
3. 저번 캠페인을 경험한 그룹이 가입률이 높았다는 점에서, 이번에 처음으로 접한 신규 잠재 고객들을 다음 캠페인에도 포함시켜야 한다는 결론을 얻을 수 있다.
4. duration을 학습 데이터에 포함할 수 없다는 상황에서, 가장 현실성 높은 고성능 모델은 SMOTE를 포함한 LogisticRegression이다.