

# 특성 공학 기반 전환율 예측 이진 분류 모델 개발

MQL 데이터 기반 B2B 영업기회 창출 예측 모델 개발

Team DASIGHT 김수림, 김영현, 황영석

# Contents

- 01** Outline
- 02** EDA
- 03** Data Preprocessing
- 04** Feature Engineering
- 05** Modeling
- 06** Conclusion

- Project 개요

MQL Data

지도 학습 기반  
ML/DL 모델 학습

고객 영업 전환  
여부 분류

고객의 행동 및 특성을 정량적으로 측정, 분석하여 고객 지수를 만들 수 있음

고객 지수는 고객을 분석하고 이해하여, 고객에게 새로운 가치를 제공함

영업 기회 전환지수는 고객 지수의 분류로, 영업 성공 가능성이 높은 고객을 선별함

MQL 고객 정보를 활용하여 영업 전환 성공여부를 예측하는 AI모델을 개발할 수 있음

따라서 MQL 고객 정보로부터 고객의 영업 전환 여부를 예측하는 이진 분류 모델 개발을 목적으로 함

## ● 데이터 소스 및 구조

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
bant_submit	customer_co	business_uni	com_reg_ver	customer_id	customer_ty	enterprise	historical_exi	id Strategic_vit	id Strategic_v	customer_jol	lead_descлер	inquiry_type	product_cate	product_sub	product_moi	customer_co	customer_po	response_csr	expected_time	ver_cus	ver_pro	ver_win_rate	ver_win_ratio	business_are	business_sub	lead_owner	isConverted	
1 /Quezon City/ AS	0.06666667	32160	End Custom Enterprise				purchasing	62 Quotation o	multi-split		/Quezon City/ entry level	LGEPH	less than 3 n	1	0	0.003079288	0.026845638	corporate / Engineering		0	TRUE							
1 /PH-00/Philippines/ AS	0.06666667	23122	End Custom Enterprise	12			media and c	96 Quotation o	multi-split		/PH-00/Philippines/ ceo/founder	LGEPH	less than 3 n	1	0	0.003079288	0.026845638	corporate / Advertising		1	TRUE							
1 /Kolkata /Inc AS	0.08888889	1755	End Custom Enterprise	144			engineering	56 Product Infra	single-split		/Kolkata /Inc partner	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / Construction		2	TRUE							
1 /Bhubanesw/ AS	0.08888889	4919	End Custom Enterprise				entrepreneur	44 Quotation o	vrf		/Bhubanesw/ ceo/founder	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / IT/Software		3	TRUE							
1 /Hyderabad/ AS	0.08888889	17126	Specifier/ Inf Enterprise				consulting	97 Quotation o	multi-split		/Hyderabad/ partner	LGEIL	less than 3 n	0	0	0.003079288	0.026845638	corporate / office		4	TRUE							
1 /Abuja/Niger AS	0.040816327	16328	End Custom SMB				program and	1114 Quotation o	chiller		/Abuja/Niger manager	LGEAF	less than 3 n	1	0	0.003079288	0.026845638	corporate / Engineering		5	TRUE							
0.75 /Jeddah, KSA/ AS	0.040816327	20664	End Custom SMB				engineering	420 Quotation o	single-split		/Jeddah, KSA/ manager	LGESJ		1	0	0.003079288	0.026845638	corporate / Engineering		6	TRUE							
1 /Guwahati/ir AS	0.08888889	17983	End Custom SMB				sales	205 Quotation o	vrf		/Guwahati/ir partner	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / Manufacturi		7	TRUE							
0.75 /Cebu city/P AS	0.06666667	30867	Specifier/ Inf Enterprise	3			other	103 Quotation o	multi-split		/Cebu city/P vice presider	LGEPH	less than 3 n	0	0	0.003079288	0.026845638	corporate / Construction		8	TRUE							
0.75 /hauz khas,d AS	0.08888889	6084	End Custom SMB				other	252 Quotation o	vrf		/hauz khas,d manager	LGEIL	3 months ~	1	0	0.003079288	0.026845638	corporate / office		9	TRUE							
0.75 /hosur/India AS	0.08888889	15379	End Custom Enterprise				engineering	90 Quotation o	vrf		/hosur/India associate/analyst	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / Manufacturi		10	TRUE							
0.75 /Koduvai, Tir AS	0.08888889	31561	End Custom SMB	23			operations	84 Quotation o	etc.		/Koduvai, Tir c-level execu	LGEIL	3 months ~	1	0	0.003079288	0.026845638	corporate / IT/Software		11	TRUE							
0.75 /Kolkata /Inc AS	0.08888889	37134	Service Part Enterprise				entrepreneur	67 Quotation o	vrf		/Kolkata /Inc partner	LGEIL	less than 3 n	0	0	0.003079288	0.026845638	corporate / Energy		3	TRUE							
0.5 /Benin City/I AS	0.040816327	30294	End Custom SMB				purchasing	210 Quotation o	vrf		/Benin City/I associate/analyst	LGEAF		1	0	0.003079288	0.026845638	corporate / Construction		12	TRUE							
0.25 /Lagos/Niger AS	0.040816327	16481	End Custom Enterprise					46 Quotation o	multi-split		/Lagos/Niger none	LGEAF		1	0	0.003079288	0.026845638	corporate / Construction		12	TRUE							
0.75 /Riyadh/Sau AS	0.040816327	22897	Specifier/ Inf SMB				engineering	166 Quotation o	chiller		/Riyadh/Sau associate/analyst	LGESJ	less than 3 n	0	0	0.003079288	0.026845638	corporate / office		13	TRUE							
0.75 /Singapore/S AS	0.06666667	1901	Specifier/ Inf SMB				engineering	129 Quotation o	vrf		/Singapore/entry level	LGESL	less than 3 n	0	0	0.003079288	0.026845638	corporate / Engineering		14	TRUE							
0.5 /Singapore/S AS	0.06666667	46362	Channel Par SMB	47			engineering	3 Quotation o	single-split		/Singapore/director	LGESL		0	0	0.003079288	0.026845638	corporate / office		14	TRUE							
1 //Philippines AS	0.06666667	28695	SMB				administrativ	57 Quotation o	single-split		//Philippines ceo/founder	LGEPH	less than 3 n	0	0	0.003079288	0.026845638	corporate / office		15	FALSE							
1 /Cebu City/P AS	0.06666667	8402	Enterprise	0			engineering	80 Quotation o	etc.		/Cebu City/P entry level	LGEPH	less than 3 n	0	0	0.003079288	0.026845638	corporate / office		16	FALSE							
0.75 /Noida/India AS	0.08888889	39910	SMB				administrativ	43 Quotation o	multi-split		/Noida/India manager	LGEIL		0	0	0.003079288	0.026845638	corporate / office		17	FALSE							
1 /BENGALURU AS	0.08888889	14295	End Custom SMB				engineering	407 Quotation o	vrf		/BENGALURU manager	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / Developer/Pla		18	FALSE							
1 /chennai/Ind AS	0.08888889	33229	End Custom Enterprise				consulting	75 Quotation o	multi-split		/chennai/Ind manager	LGEIL	less than 3 n	1	0	0.003079288	0.026845638	corporate / Entertainmen		19	FALSE							
1 /ghaziabad/I AS	0.08888889	45122	SMB				purchasing	66 Quotation o	vrf		/ghaziabad/I manager	LGEIL	less than 3 n	0	0	0.003079288	0.026845638	corporate / office		20	FALSE							
0.75 /Pune/India AS	0.08888889	39810	End Custom SMB				purchasing	71 Quotation o	etc.		/Pune/India other	LGEIL		1	0	0.003079288	0.026845638	corporate / Manufacturi		21	FALSE							
1 /KANPUR/In AS	0.08888889	20437	Specifier/ Inf Enterprise	1			engineering	113 Quotation o	rac		/KANPUR/In partner	LGEIL	less than 3 n	0	0	0.003079288	0.026845638	corporate / Engineering		22	FALSE							
1 /RS/Brazil AS	0.003937008	12109	Enterprise				other	133 Quotation o	teto ou cassette inverter		/RS/Brazil other	LGESP	3 months ~	0	0	0.003079288	0.026845638	corporate / office		23	FALSE							
1 /Dubai/U.A.E AS	0.040816327	44685	End Custom Enterprise				purchasing	142 Quotation o	vrf		/Dubai/U.A.E manager	LGEFF	less than 3 n	1	0	0.003079288	0.026845638	corporate / Construction		24	FALSE							
1 /Dubai/U.A.E AS	0.040816327	4504	End Custom SMB				purchasing	76 Quotation o	vrf		/Dubai/U.A.E manager	LGEFF	less than 3 n	1	0	0.003079288	0.026845638	corporate / Construction		25	FALSE							
1 /Dubai/U.A.E AS	0.040816327	29418	Specifier/ Inf Enterprise				engineering	334 Quotation o	vrf		/Dubai/U.A.E vice presider	LGEFF	less than 3 n	0	0	0.003079288	0.026845638	corporate / Construction		25	FALSE							

## MQL(Marketing Qualified Lead)

특정 기준을 충족하여 판매 가능성이 높은 것으로 평가된  
마케팅 자격을 갖춘 잠재 고객의 정보가 기록된 데이터

## ● 데이터 소스 및 구조

The image shows two side-by-side forms from the LG Business Solutions website.

**Inquiry For Integrated Solutions (Left):**

- Select Products (At least two):**
  - Monitor Signage
  - Commercial TV
  - ESS
  - Monior/Monitor TV
  - PC
  - Projector
  - Robot
  - System AC
  - EMS
  - RAC
  - Chiller
  - TV
  - Refrigerator
  - Washing Machine
  - Aircare
  - Vacuum Cleaner
  - Styler
  - Dryer
  - Built-in/Cooking
  - Home Beauty
  - Water Care
  - Audio/Video
- Inquiry Type \***: Inquiry Type dropdown menu.
- Timeline \***: Timeline dropdown menu.
- Budget \***: Budget dropdown menu.
- Message \***: Large text area for message.

**Personal Information (Right):**

- Region \***: Region dropdown menu.
- Country \***: Country dropdown menu.
- State**: State dropdown menu.
- Province/City**: Text input field.
- First Name \***: First Name text input field.
- Last Name \***: Last Name text input field.
- Phone Number**: Phone Number text input field.
- Work E-Mail \***: Work E-Mail text input field.
- Customer Type \***: Customer Type dropdown menu.
- Customer Sub Type**: Customer Sub Type dropdown menu.
- Company Name \***: Company Name text input field.
- Business Sector (Lv1)**: Business Sector (Lv1) dropdown menu.
- Business Sector (Lv2)**: Business Sector (Lv2) dropdown menu.
- Job Function \***: Job Function dropdown menu.
- Seniority Level \***: Seniority Level dropdown menu.

고객이 직접 작성한 관심 상품과 본인 정보 + 영업 성공 여부(이진 라벨)  
BANT 요소, 고객 및 사업 관련 정보, 영업 전환 이력 등 포함

- 데이터 소스 및 구조

### 고객 정보

bant\_submit  
customer\_country  
business\_unit  
com\_reg\_ver\_win\_rate  
customer\_idx  
customer\_type  
enterprise  
historical\_existing\_cnt  
id\_strategic\_ver  
it\_strategic\_ver  
idit\_strategic\_ver  
customer\_job  
. . .



영업 전환 성공 여부 '**isConverted**' 예측

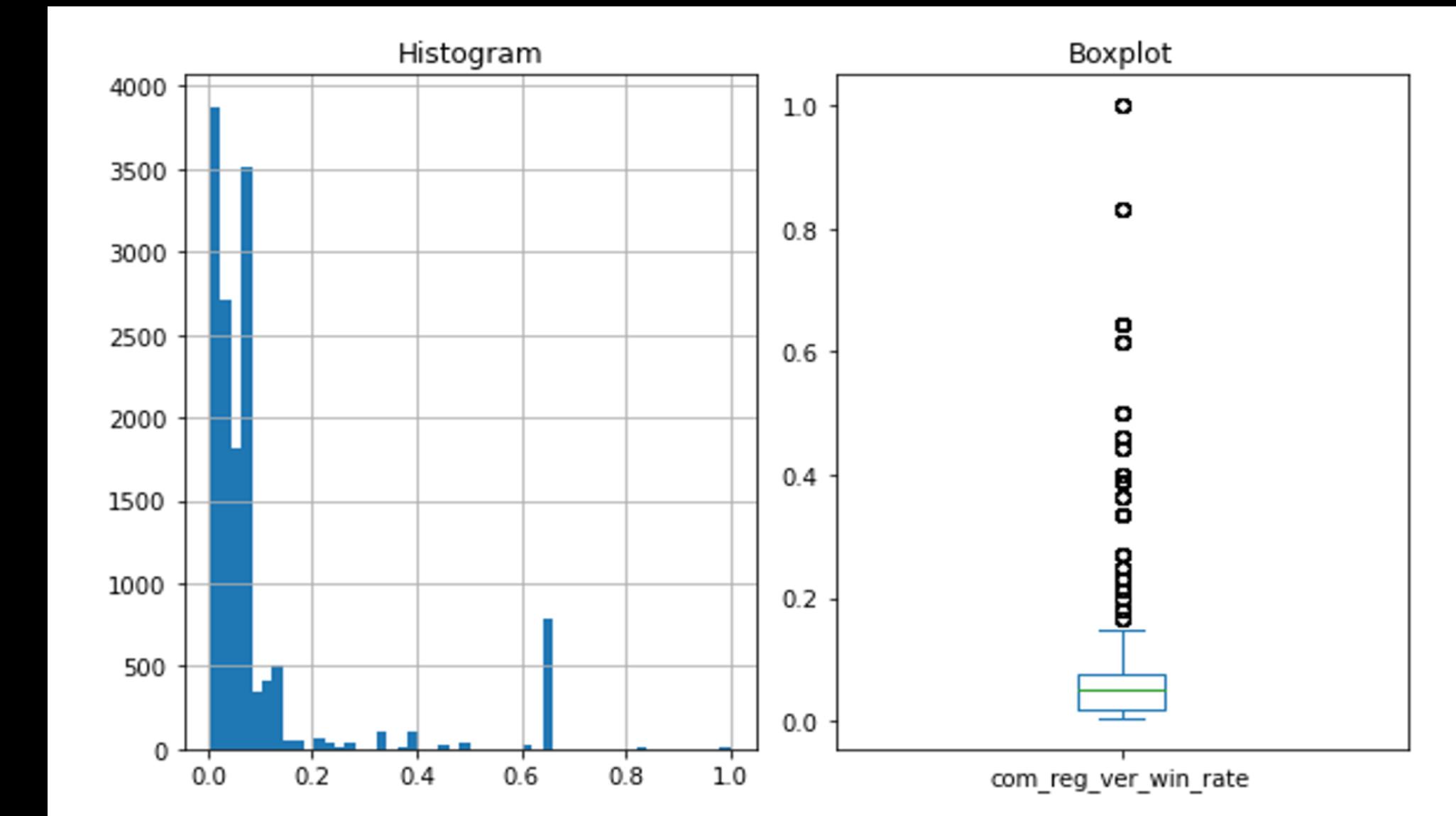
28개의 Columns

- 통계적 요약 - 수치형 변수

### com\_reg\_ver\_win\_rate

Vertical Level 1, business unit, region을 기준으로 oppy 비율을 계산  
대부분의 값이 0에 가까움, 중앙값이 낮으며 이상치 존재

Info	Data
데이터 개수	14568개
결측치 개수	44731개 (75.46%)
고유값 개수	81개
Mean	0.091685
Std	0.150988
Min	0.003788
25%	0.019900
50%	0.049180
75%	0.074949
Max	1.000000



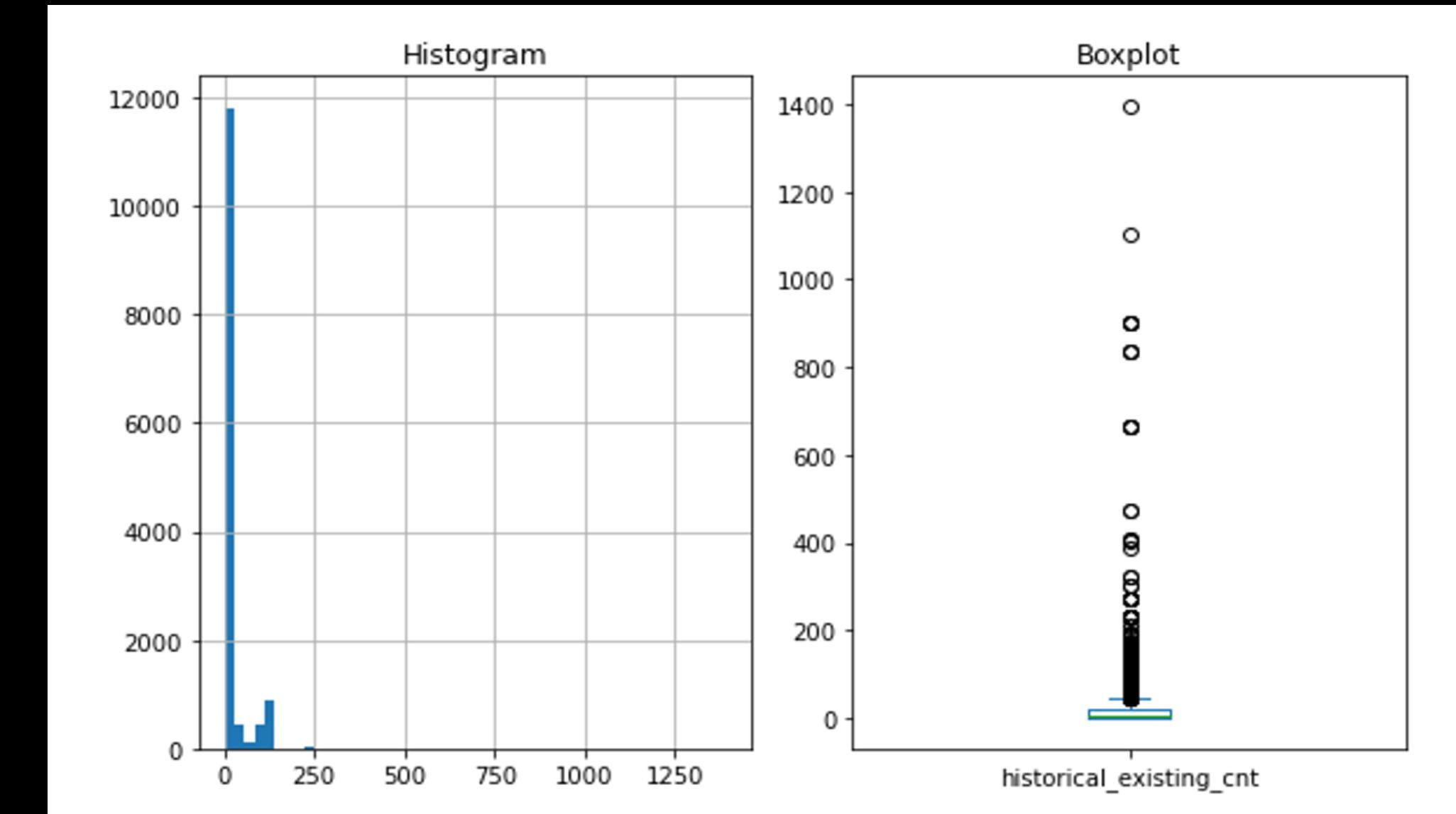
- 통계적 요약 - 수치형 변수

### historical\_existing\_cnt

이전에 Converted(영업 전환) 되었던 횟수

대부분의 값이 0에 가까움, 중앙값이 낮으며 이상치 존재

Info	Data
데이터 개수	13756개
결측치 개수	45543개 (76.82%)
고유값 개수	137개
Mean	19.912184
Std	44.697938
Min	0.000000
25%	1.000000
50%	4.000000
75%	19.000000
Max	1394.000000

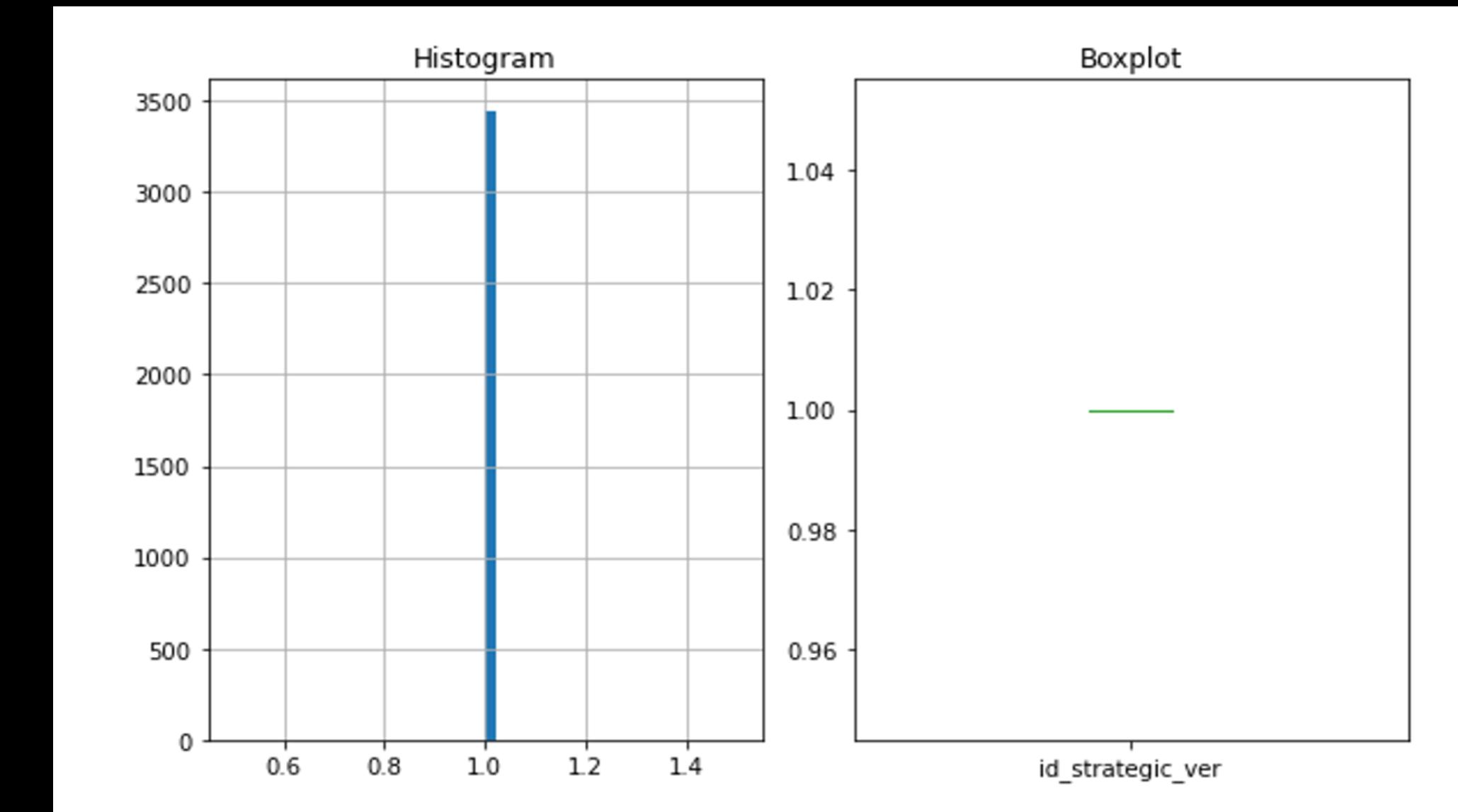


- 통계적 요약 - 수치형 변수

**id\_strategic\_ver**

(도메인 지식) 특정 사업부(Business Unit), 특정 사업 영역(Vertical Level1)에 대해 가중치를 부여, 모두 1의 값을 가지고 있으며, 많은 결측치 존재

Info	Data
데이터 개수	3440개
결측치 개수	55855개 (94.19%)
고유값 개수	2개
Mean	1.000000
Std	0.000000
Min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
Max	1.000000

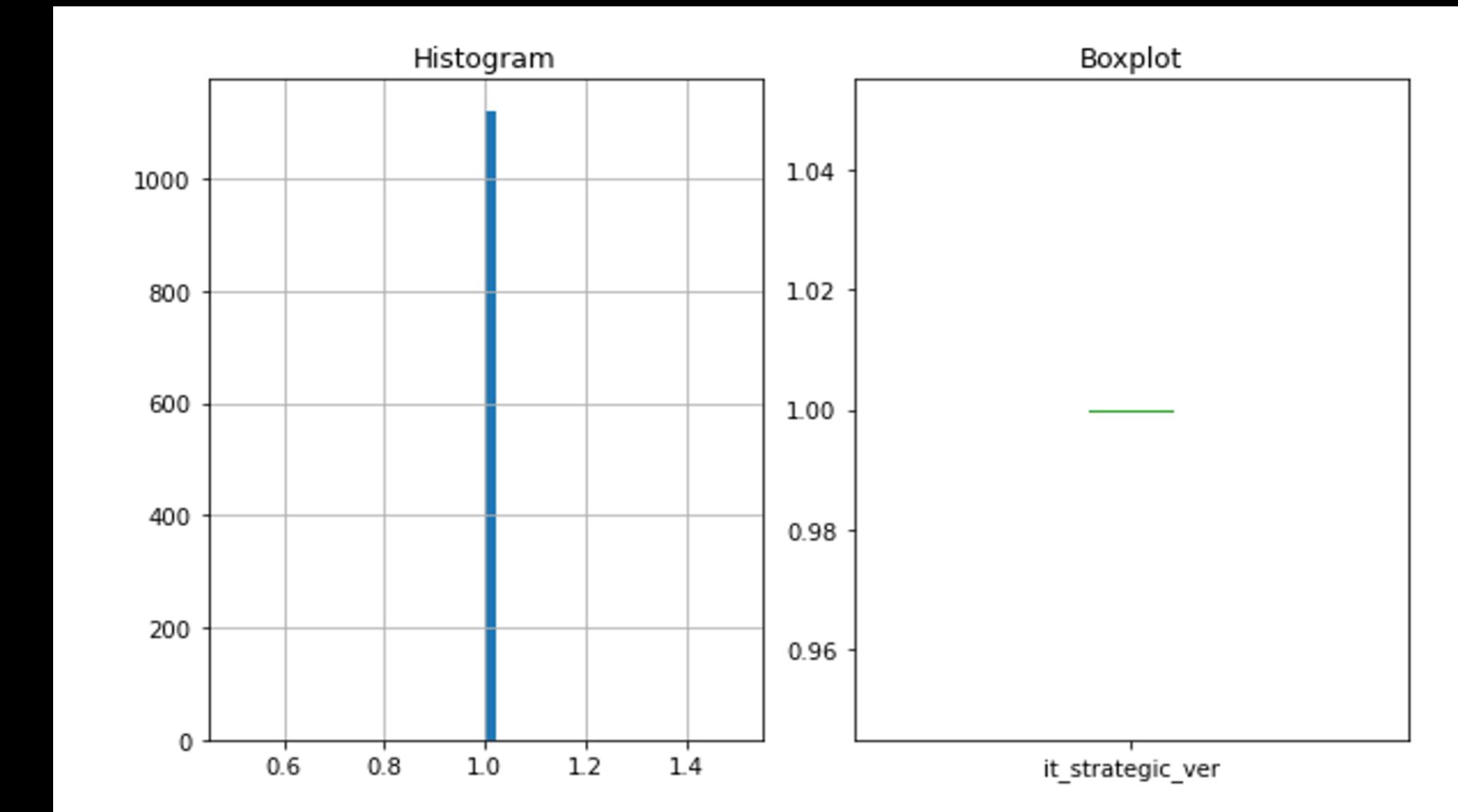


- 통계적 요약 - 수치형 변수

it\_strategic\_ver

(도메인 지식) 특정 사업부(Business Unit), 특정 사업 영역(Vertical Level1)에 대해 가중치를 부여, 모두 1의 값을 가지고 있으며, 많은 결측치 존재

Info	Data
데이터 개수	3440개
결측치 개수	58178개 (98.11%)
고유값 개수	2개
Mean	1.000000
Std	0.000000
Min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
Max	1.000000

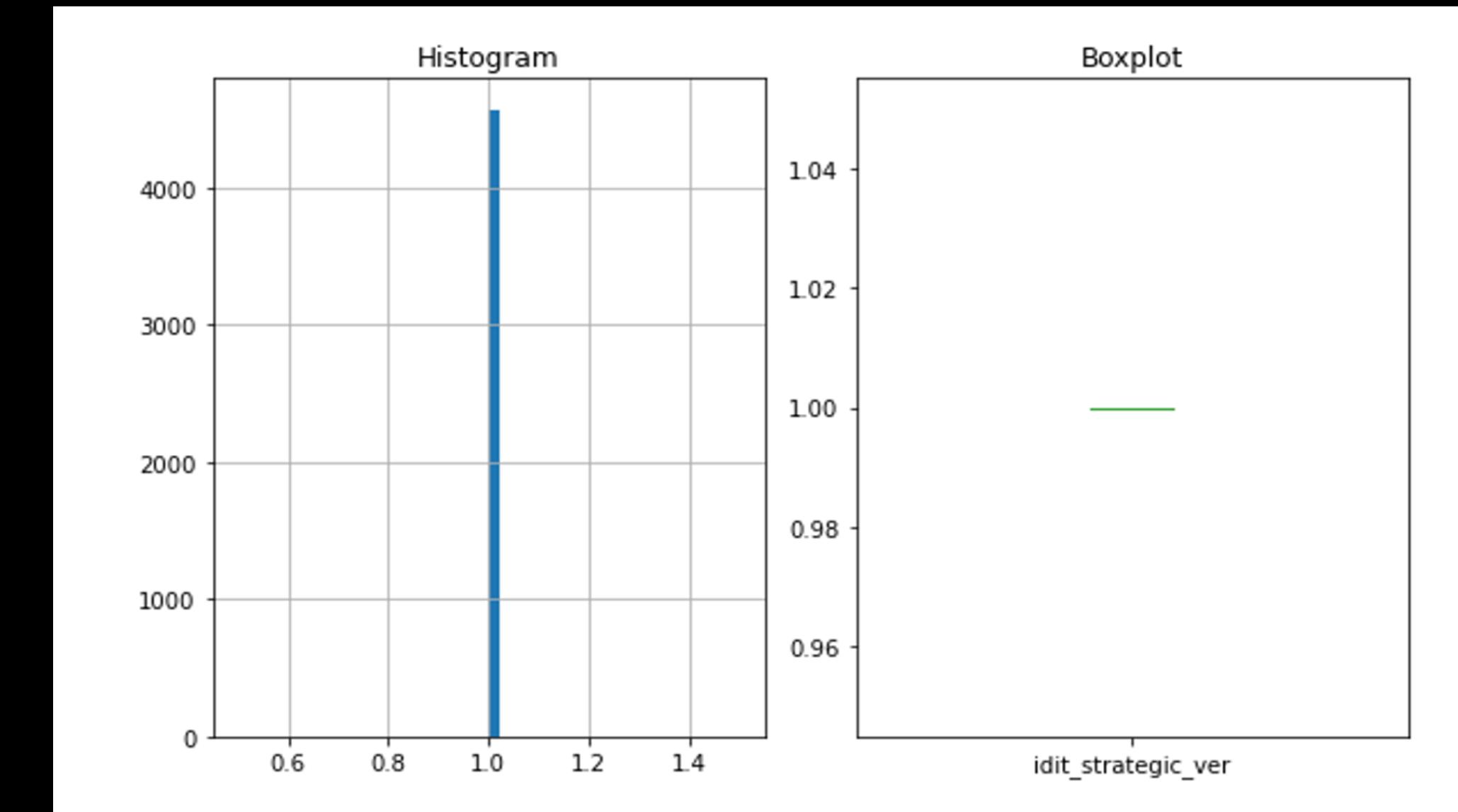


- 통계적 요약 - 수치형 변수

### idit\_strategic\_ver

Id\_strategic\_ver이나 it\_strategic\_ver 값 중 하나라도 1의 값을 가지면 1 값으로 표현  
모두 1의 값을 가지고 있으며, 많은 결측치 존재

Info	Data
데이터 개수	4565개
결측치 개수	54734개 (92.31%)
고유값 개수	2개
Mean	1.000000
Std	0.000000
Min	1.000000
25%	1.000000
50%	1.000000
75%	1.000000
Max	1.000000

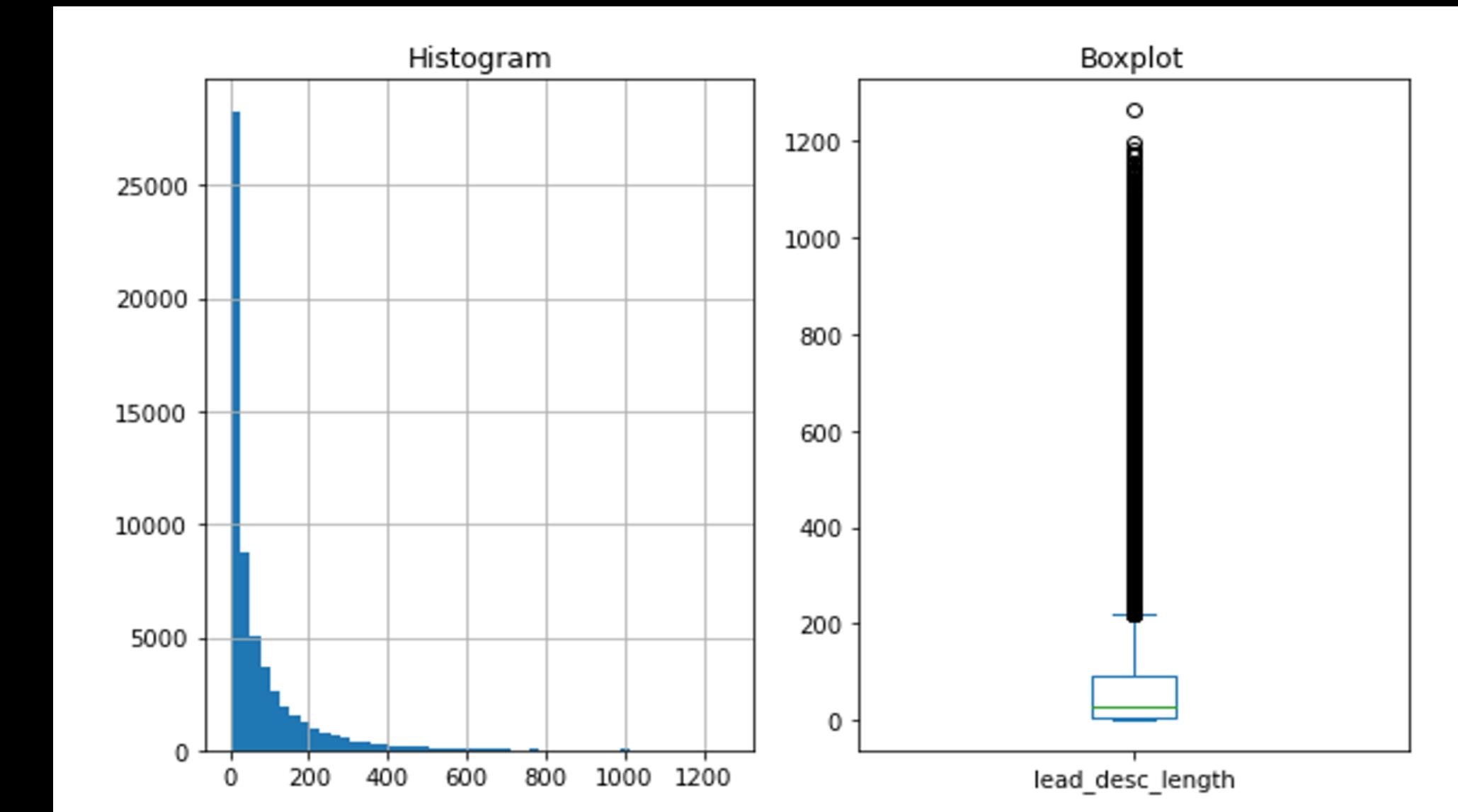


- 통계적 요약 - 수치형 변수

### lead\_desc\_length

고객이 작성한 Lead Description 텍스트 총 길이  
많은 이상치들이 존재, 설명 길이가 매우 다양함

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	909개
Mean	79.271590
Std	132.551067
Min	1.000000
25%	7.000000
50%	29.000000
75%	92.000000
Max	1264.000000

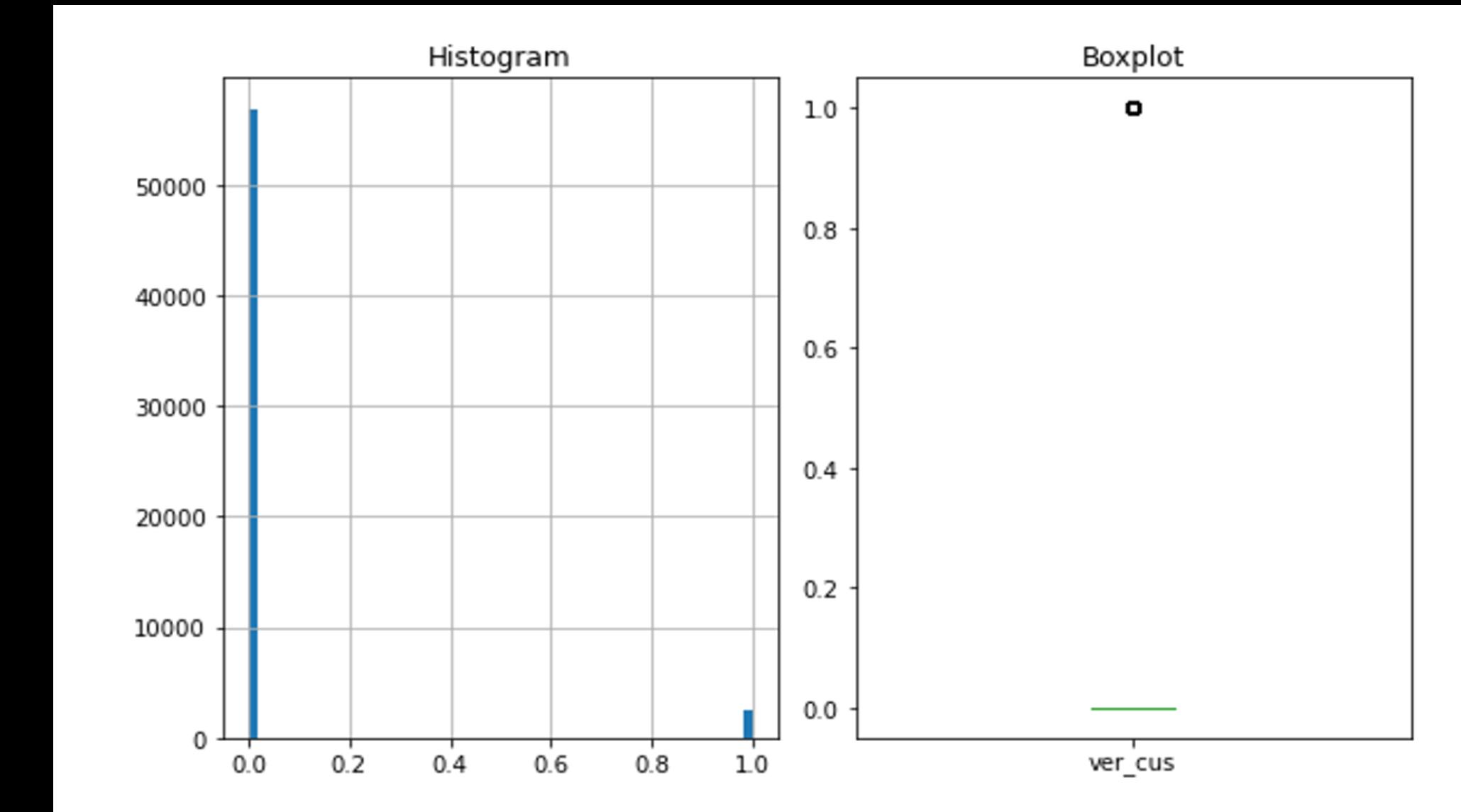


- 통계적 요약 - 수치형 변수

ver\_cus

특정 Vertical Level 1(사업영역) 이면서 Customer\_type(고객 유형)이 소비자(End-user)인 경우에 대한 가중치, 대부분의 데이터가 0에 집중되어 있음

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	2개
Mean	0.041603
Std	0.199681
Min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
Max	1.000000

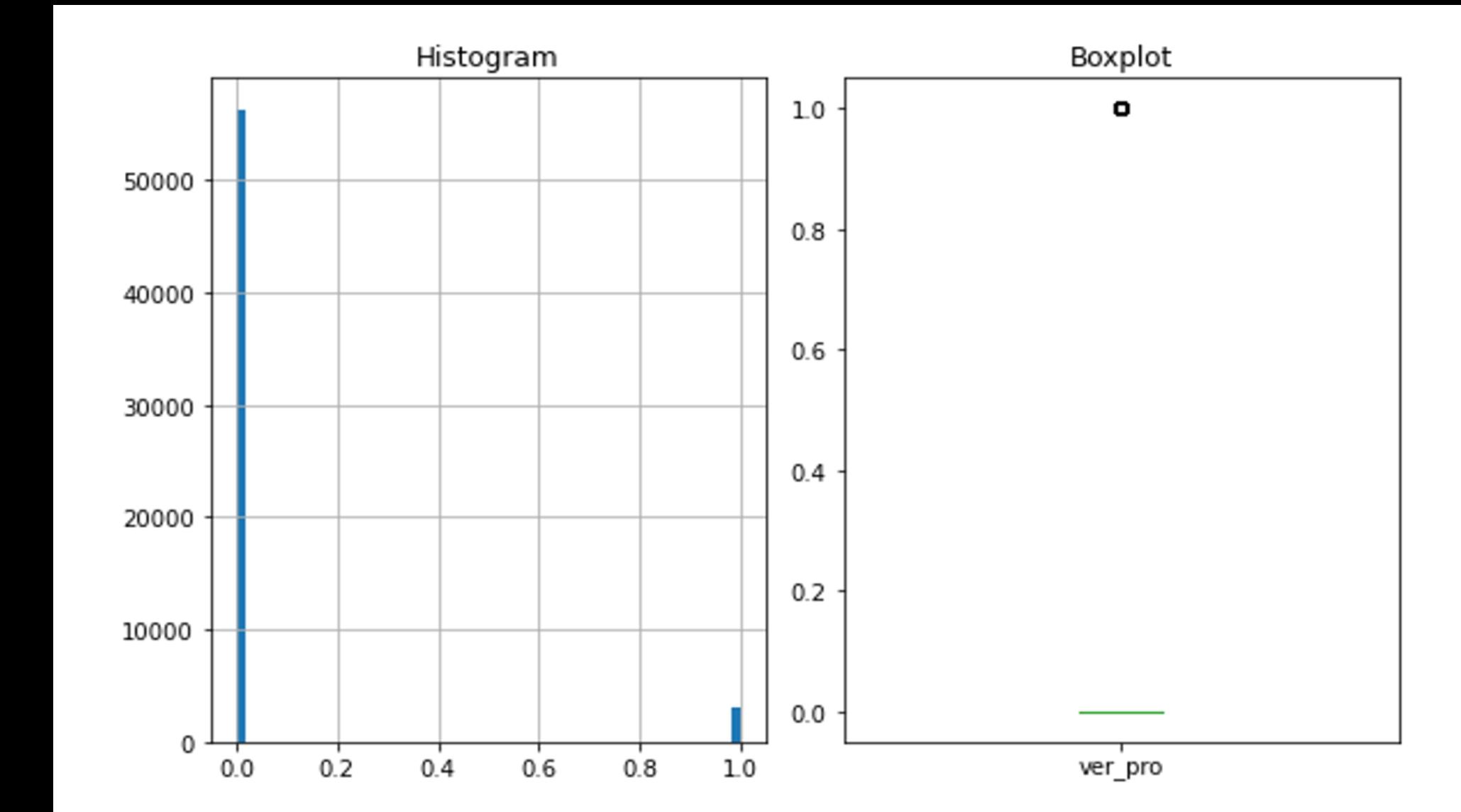


- 통계적 요약 - 수치형 변수

ver\_pro

특정 Vertical Level 1(사업영역) 이면서 특정 Product Category(제품 유형)인 경우에 대한  
가중치, 대부분의 데이터가 0에 집중되어 있음

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	2개
Mean	0.050810
Std	0.219612
Min	0.000000
25%	0.000000
50%	0.000000
75%	0.000000
Max	1.000000

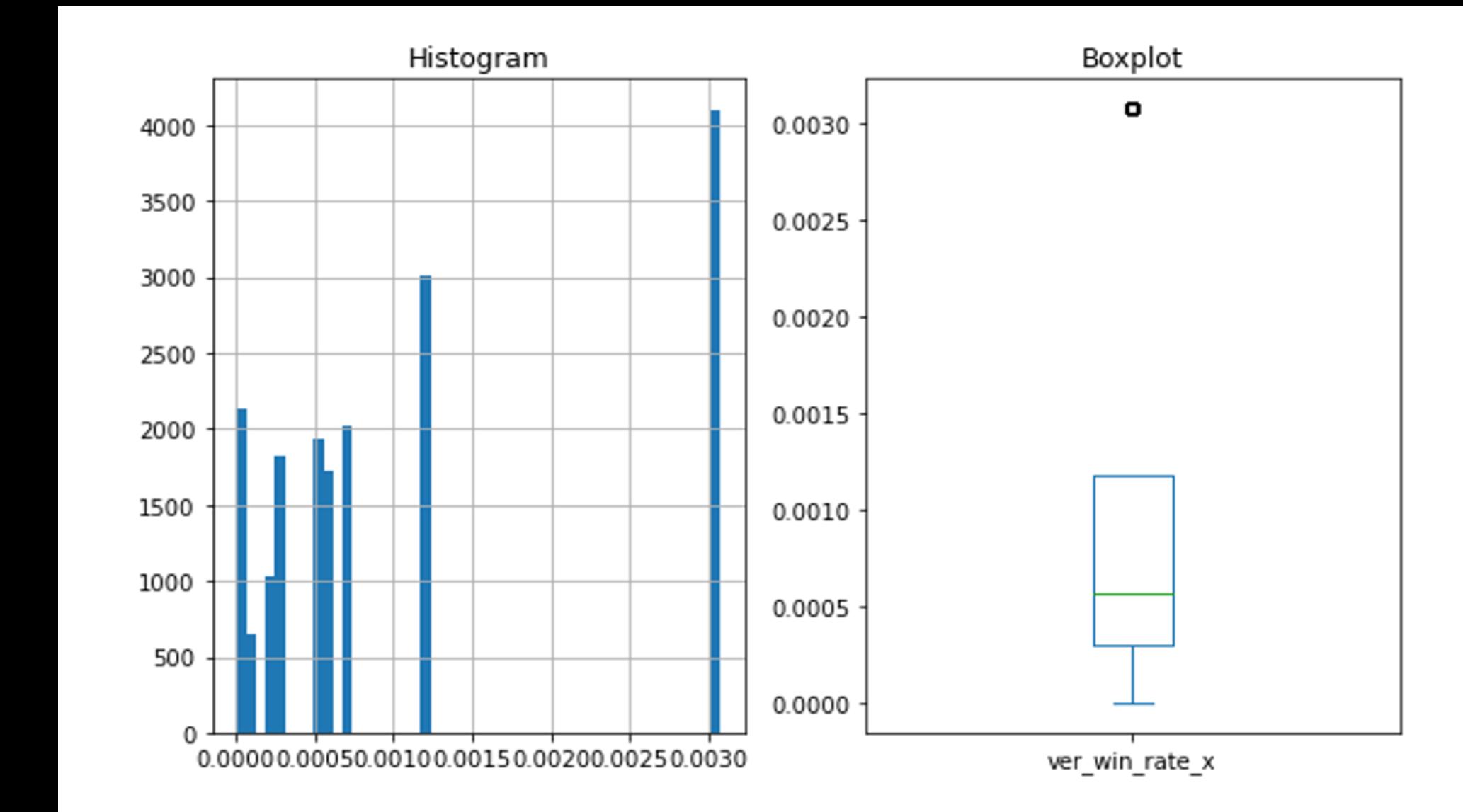


- 통계적 요약 - 수치형 변수

### ver\_win\_rate\_x

전체 Lead 중에서 Vertical을 기준으로 Vertical 수 비율과 Vertical 별 Lead 수 대비 영업 전환 성공 비율 값을 곱한 값, 매우 작은 범위의 값에 집중되어 있음, 특정 이상치가 존재

Info	Data
데이터 개수	18417개
결측치 개수	40882개 (68.92%)
고유값 개수	13개
Mean	0.001117
Std	0.001104
Min	0.000002
25%	0.000298
50%	0.000572
75%	0.001183
Max	0.003079

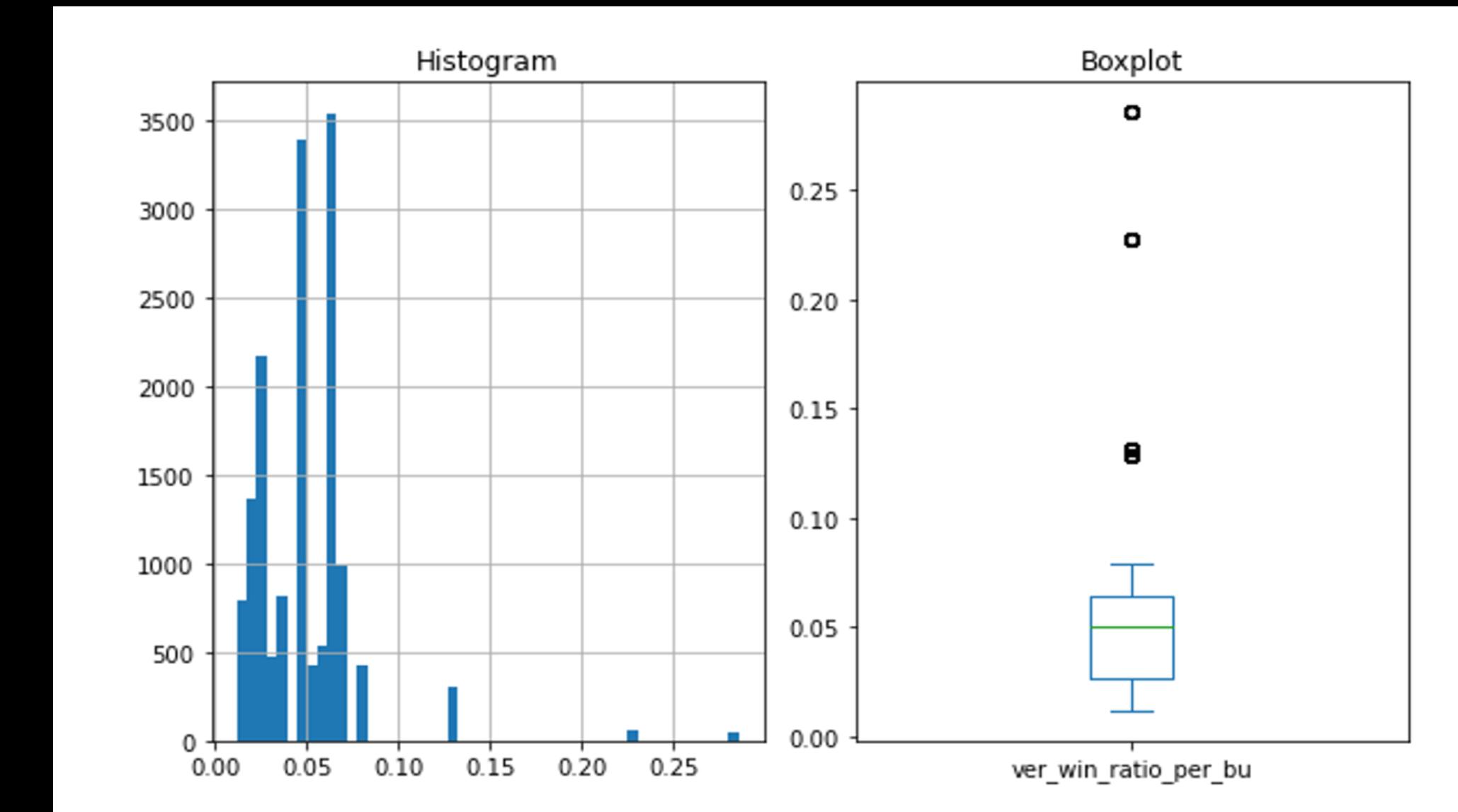


- 통계적 요약 - 수치형 변수

### ver\_win\_ratio\_per\_bu

특정 Vertical Level1의 Business Unit 별 샘플 수 대비 영업 전환된 샘플 수의 비율을 계산  
대부분 낮은 값이지만 눈에 띠는 이상치가 존재함, 특정 유닛에서 높은 승률을 암시

Info	Data
데이터 개수	15304개
결측치 개수	43995개 (74.17%)
고유값 개수	24개
Mean	0.049288
Std	0.027949
Min	0.011583
25%	0.026846
50%	0.049840
75%	0.064566
Max	0.285714



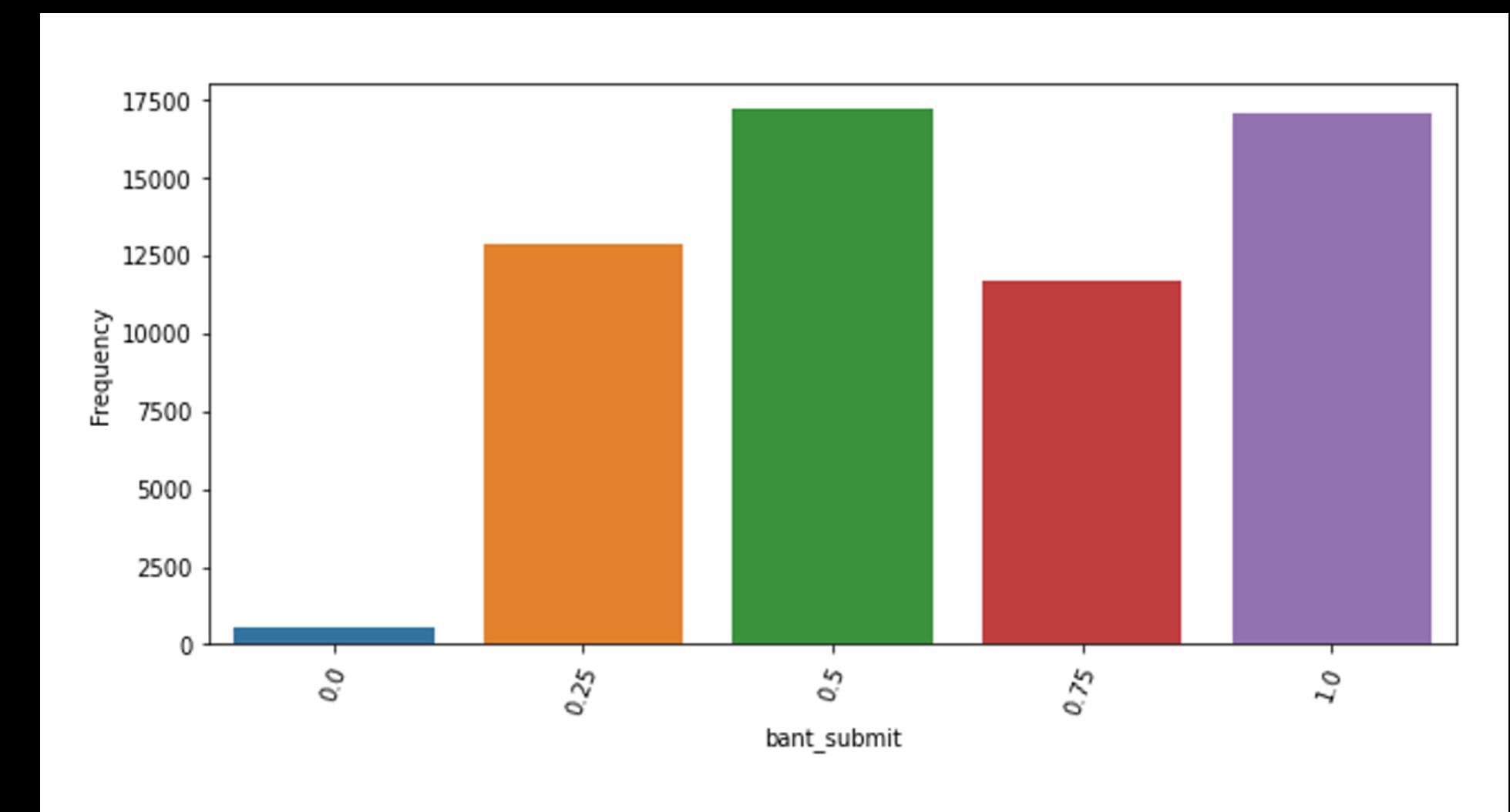
- 통계적 요약 - 범주형 변수

### bant\_submit

MQL 구성 요소들 중 [1]Budget(예산), [2]Title(고객의 직책/직급), [3]Needs(요구사항), [4]Timeline(희망 납기일) 4가지 항목에 대해서 작성된 값의 비율  
타입이 float64이지만 한정적인 고유값에 따라 범주형으로 취급

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	5개

[0, 0.25, 0.5, 0.75, 1.0]



- 통계적 요약 - 범주형 변수

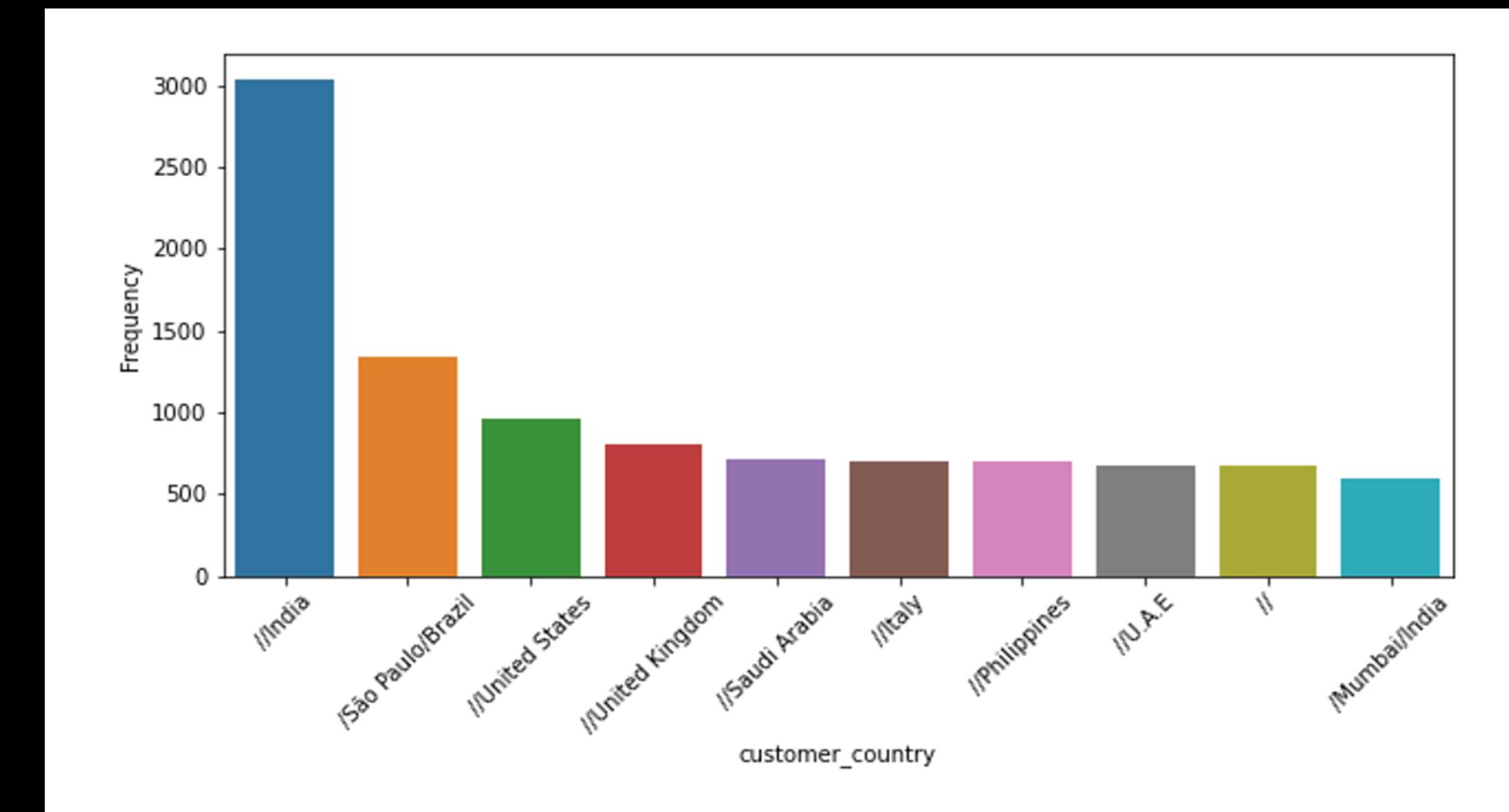
**customer\_country**

고객의 국적

/(슬래시)로 나라, 도시, 세부 주소로 구분되어 있음, 분리 작업 필요

Info	Data
데이터 개수	58317개
결측치 개수	982개 (1.65%)
고유값 개수	15400개

```
['/Quezon City/Philippines' '/PH-00/Philippines'
 '/Kolkata /India' ... '/Pisco/Peru' '/santa cruz
 bolivia/Peru' '/paris/France']
```

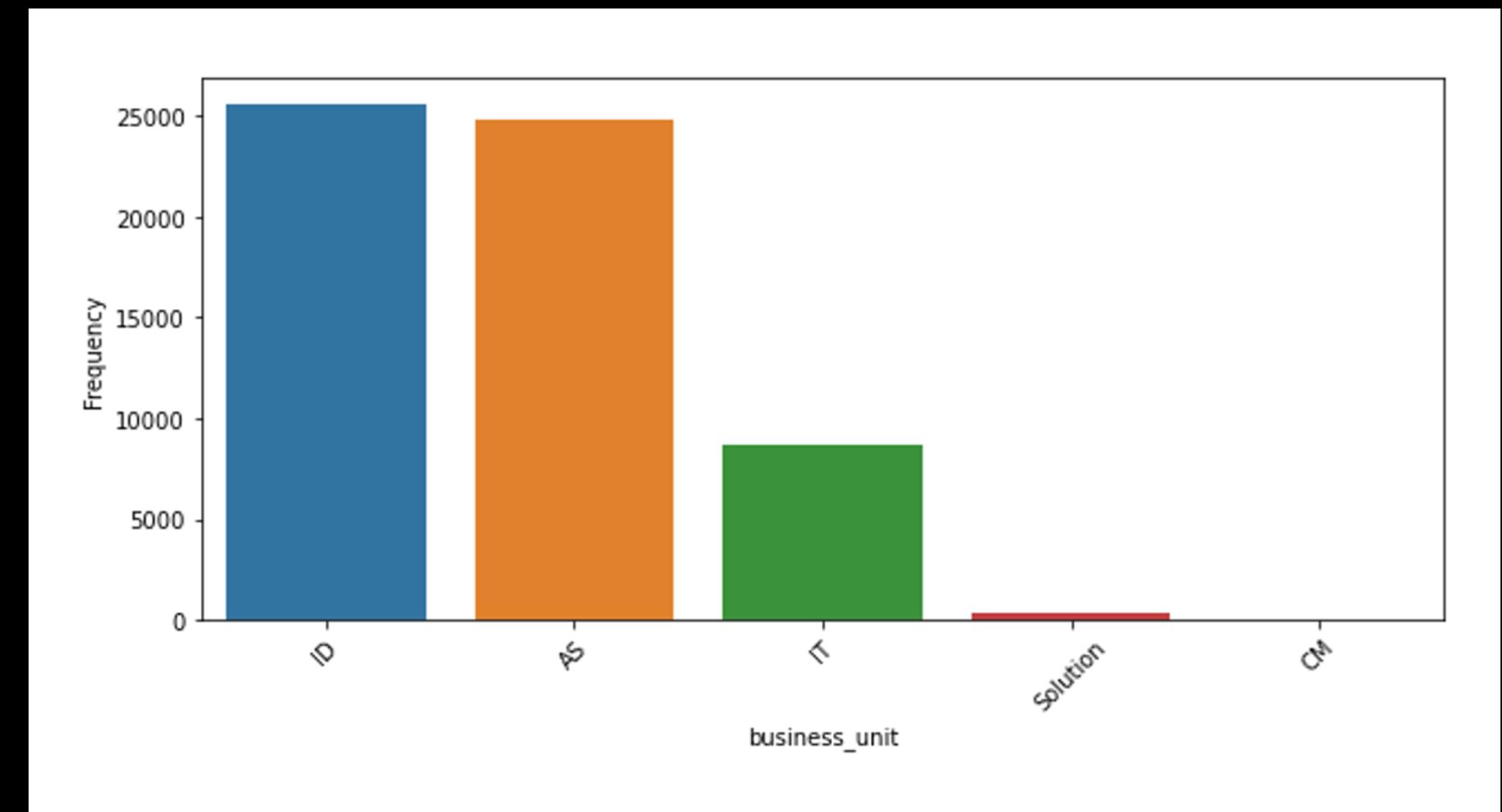


- 통계적 요약 - 범주형 변수

**business\_unit**  
MQL 요청 상품에 대응되는 사업부

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	5개

['AS' 'ID' 'IT' 'Solution' 'CM']



- 통계적 요약 - 범주형 변수

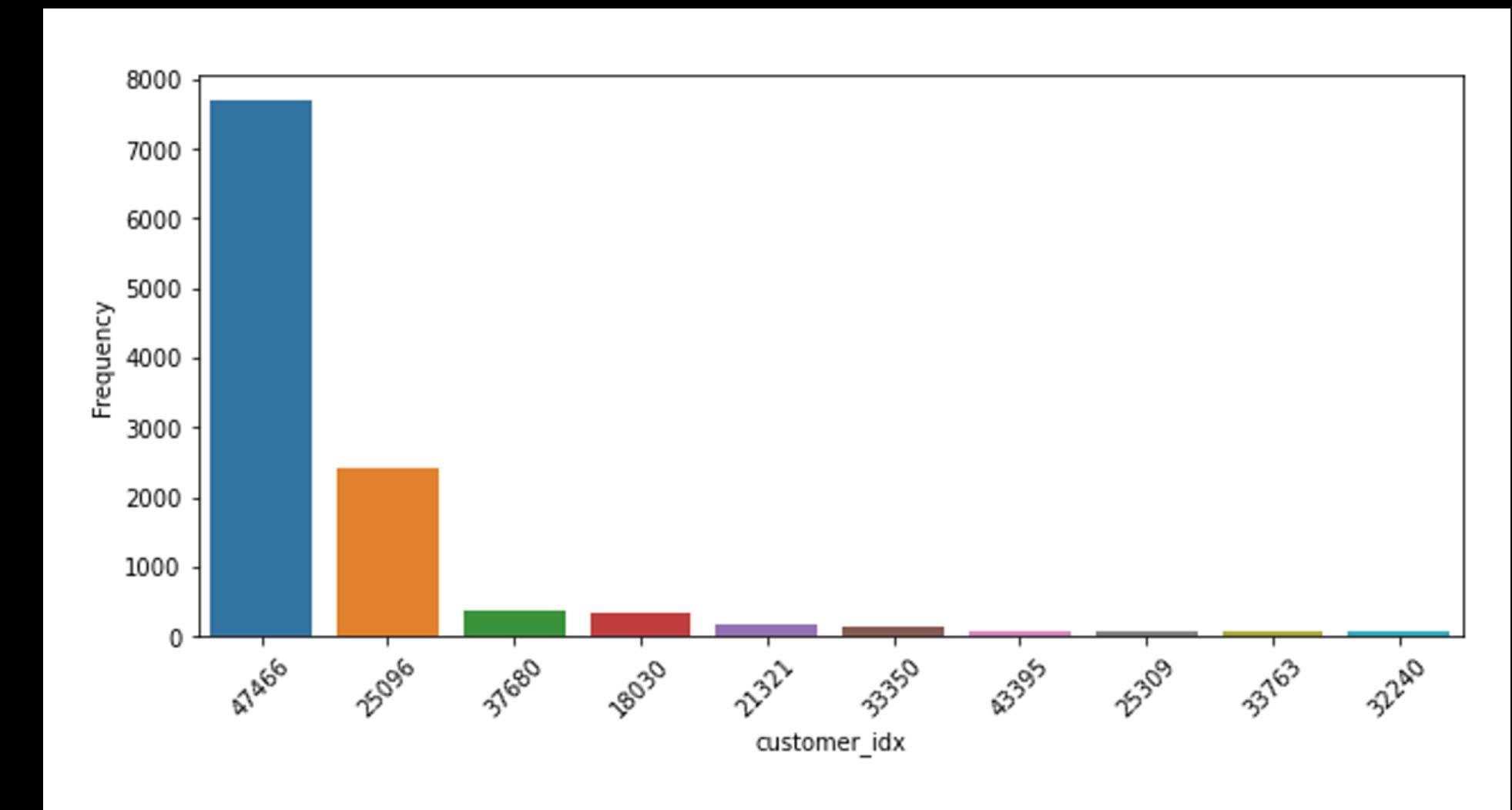
customer\_idx

고객의 회사명

타입이 int64이지만 의미에 따라 범주형으로 취급

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	35112개

[32160 23122 1755 ... 19249 40327 30268]



- 통계적 요약 - 범주형 변수

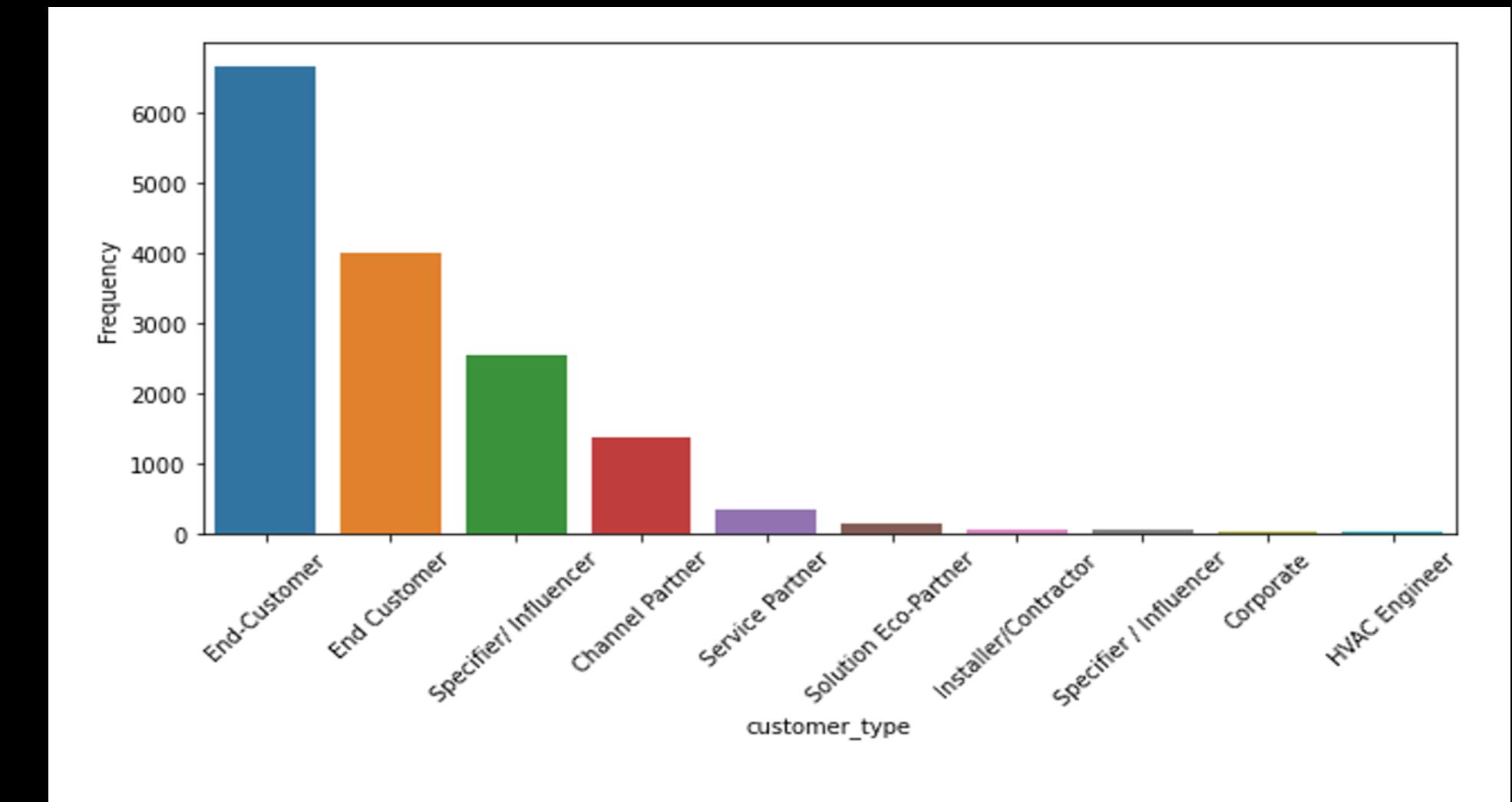
**customer\_type**

고객 유형

같은 의미지만 다르게 명시된 데이터 존재

Info	Data
데이터 개수	15338개
결측치 개수	43961개 (74.12%)
고유값 개수	34개

['End-Customer' 'Specifier/ Influencer'  
 'Service Partner' 'Channel Partner' nan  
 'Corporate' 'End Customer' ...  
 'Commercial end-user' 'Interior Designer'  
 'Home Owner' 'Administrator']

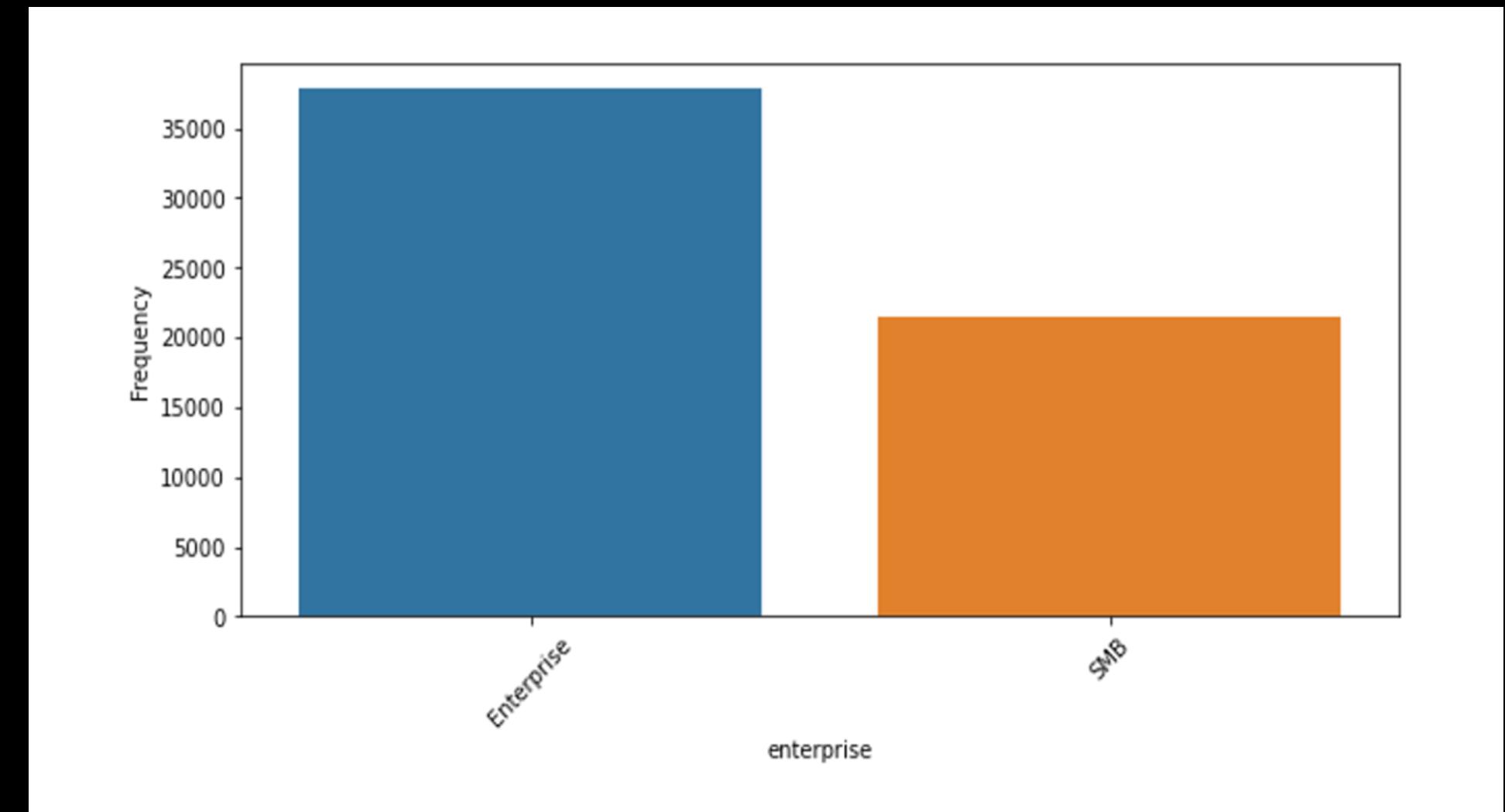


- 통계적 요약 - 범주형 변수

enterprise  
Global 기업인지, Small/Medium 규모의 기업인지

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	2개

['Enterprise' 'SMB']

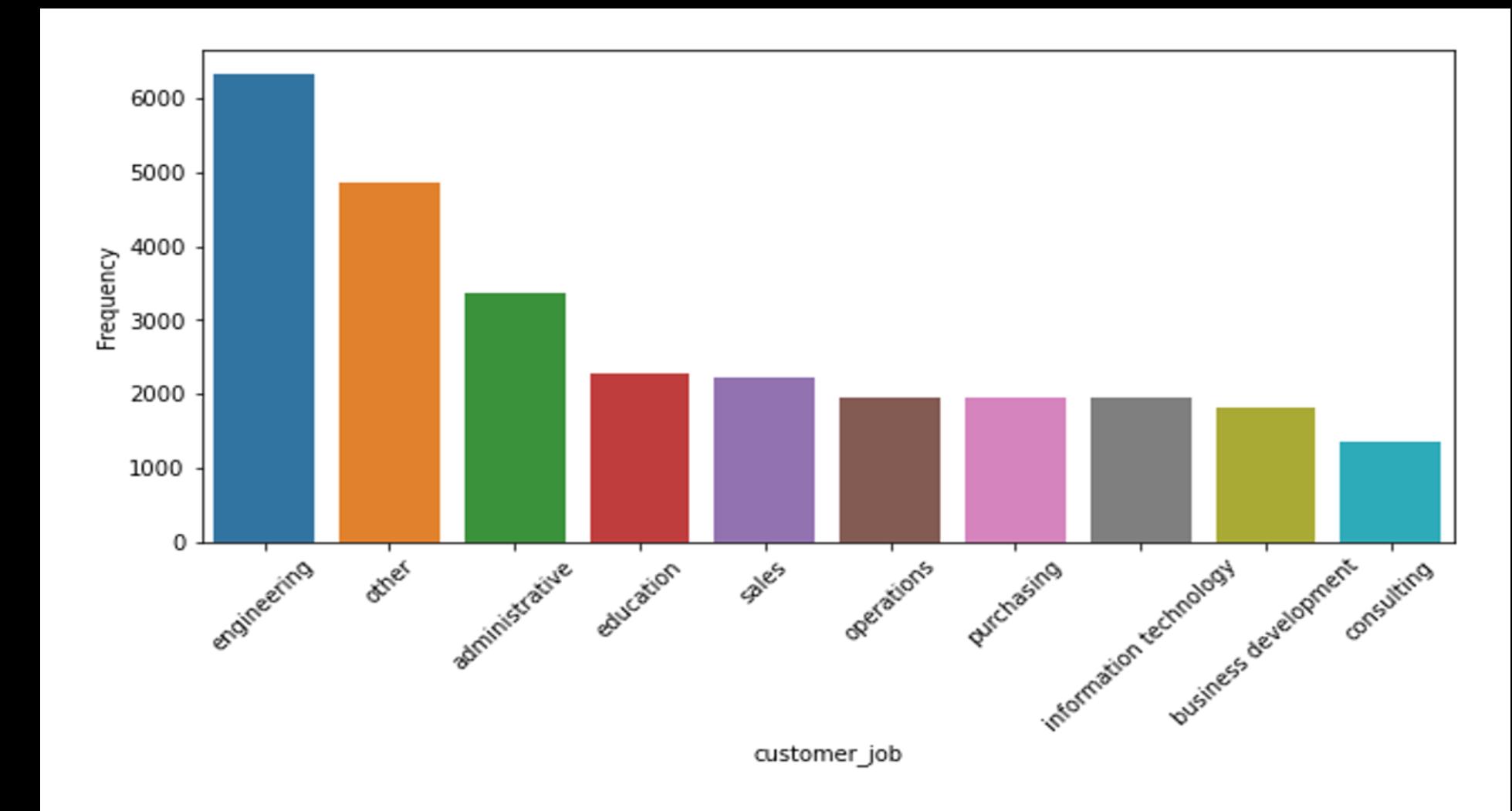


- 통계적 요약 - 범주형 변수

## customer\_job 고객의 직업군

Info	Data
데이터 개수	40566개
결측치 개수	18733개 (31.59%)
고유값 개수	561개

['purchasing' 'media and communication' 'engineering'  
'entrepreneurship' … 'developer/property' 'radiology'  
'professional' 'graphic/color art' 'medical imaging'  
'specialist' 'medical solution provider' 'manager']

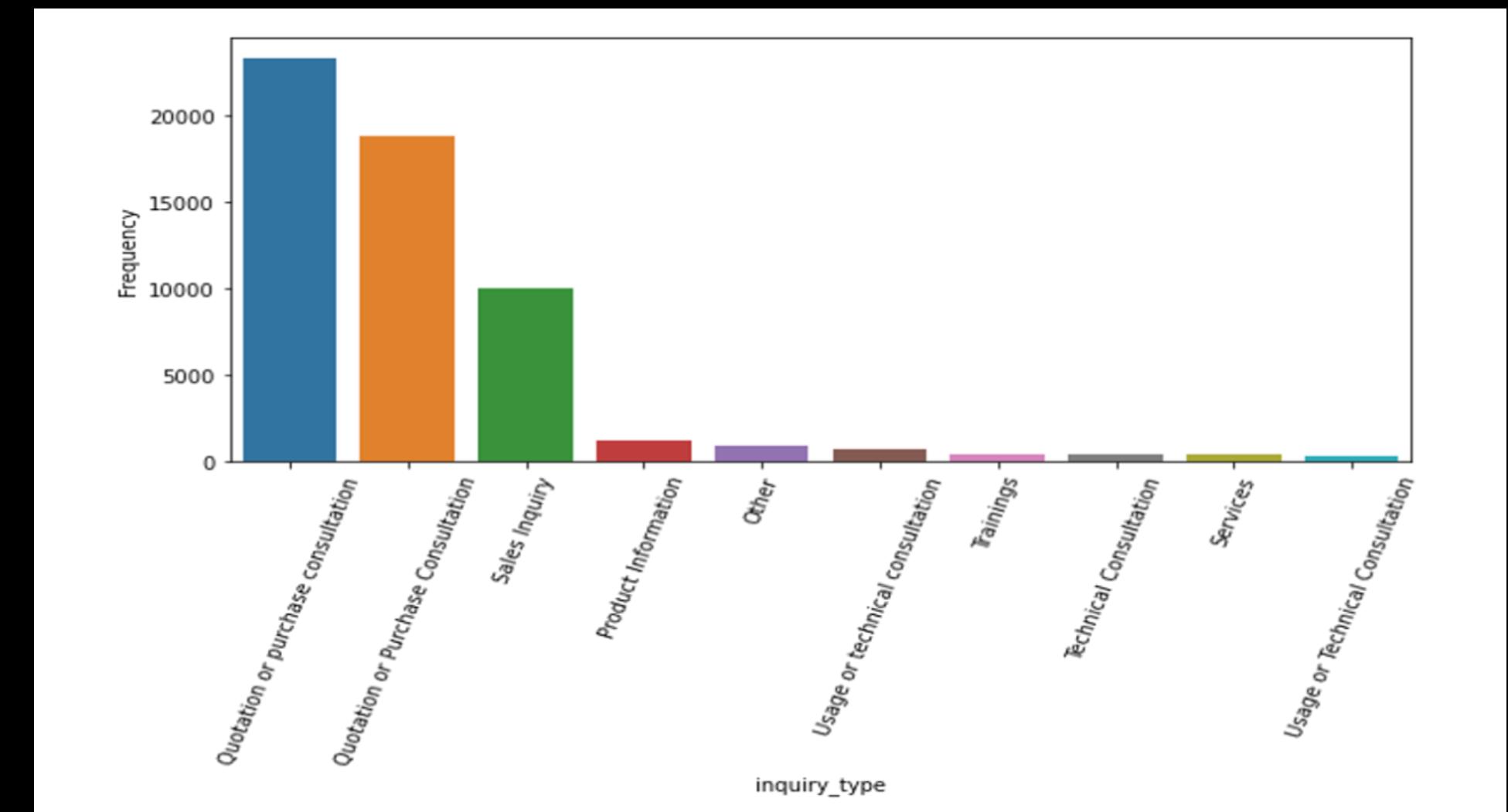


- 통계적 요약 - 범주형 변수

**inquiry\_type**  
고객의 문의 유형

Info	Data
데이터 개수	58358개
결측치 개수	941개 (1.59%)
고유값 개수	72개

[ 'Quotation or purchase consultation' 'Product Information' 'Quotation or Purchase Consultation' 'Other' 'Etc.' … 'Technical Support' 'Usage or Technical Consultation' 'Technical Consultation' 'Request for Partnership' nan 'sales' 'technical' ]

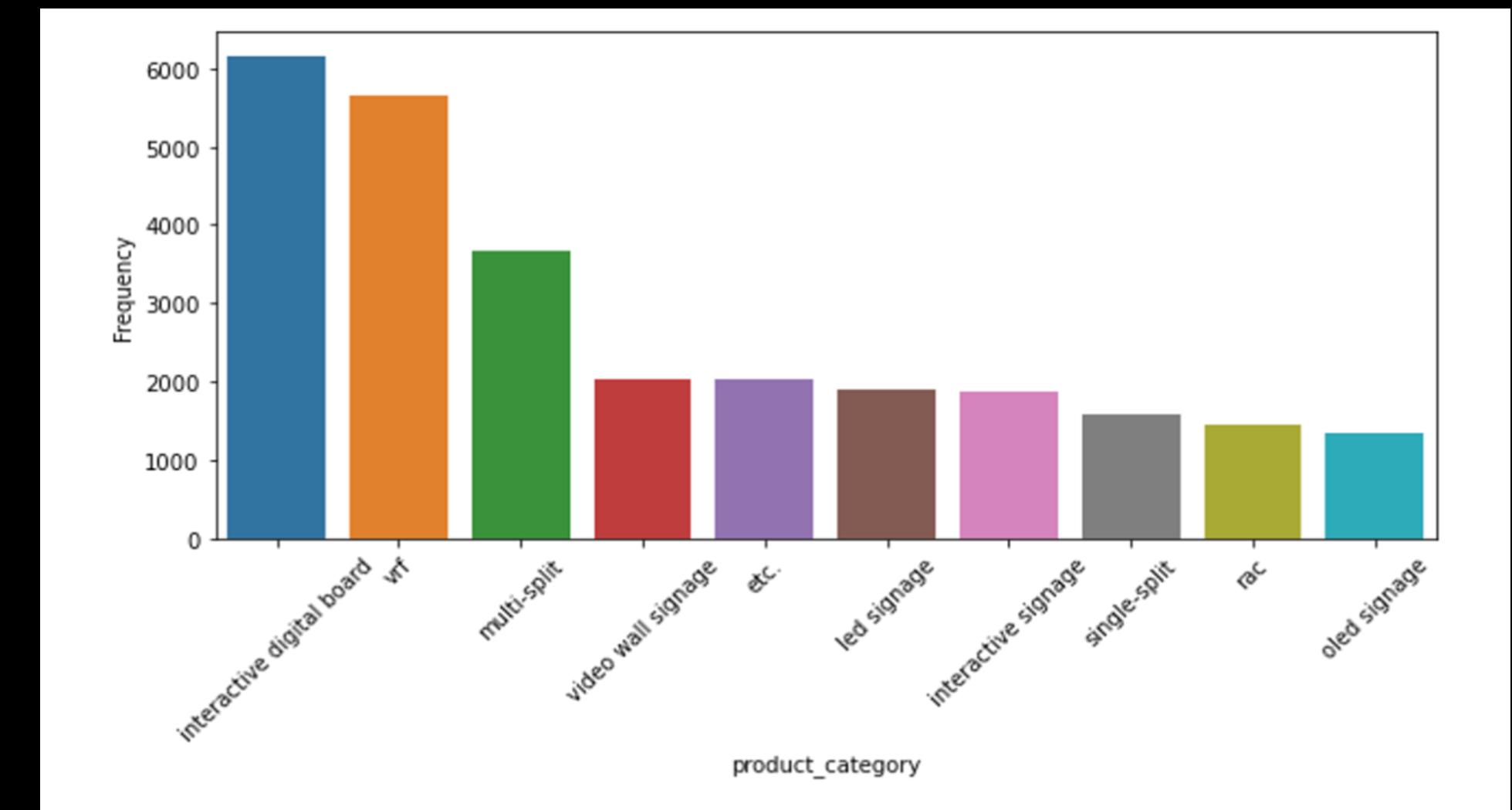


- 통계적 요약 - 범주형 변수

## product\_category 요청 제품 카테고리

Info	Data
데이터 개수	39925개
결측치 개수	19374개 (32.67%)
고유값 개수	358개

['multi-split' 'single-split' 'vrf' 'chiller' 'etc.' 'rac' 'teto  
ou cassette inverter' nan 'software solution' … 'outros'  
'heating' 'multi v5 vrf' 'split tunggal' 'multi inverter'  
'high' 'standard signage']

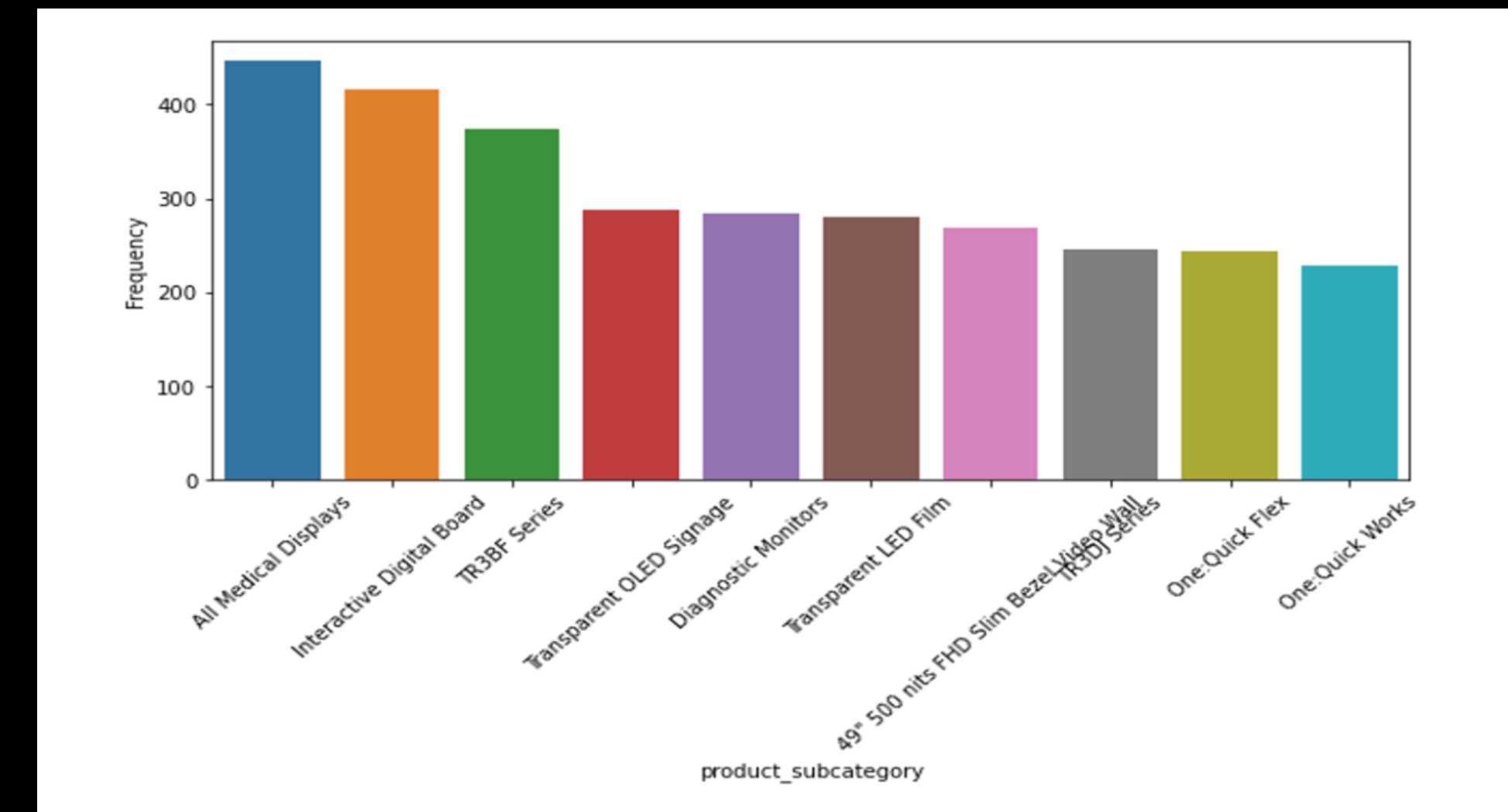


- 통계적 요약 - 범주형 변수

**product\_subcategory**  
요청 제품 하위 카테고리

Info	Data
데이터 개수	9235개
결측치 개수	50064개 (84.42%)
고유값 개수	331개

[nan 'New High Haze UHD Standard Signage' 'Window Facing Display' 'LG CreateBoard' … 'SM3G Series' '55" 500 nits FHD 0.44mm Even Bezel Video Wall' 'UH5F-H Series' 'Interactive Digital Board' 'Createboard' 'UHD TV Signage']

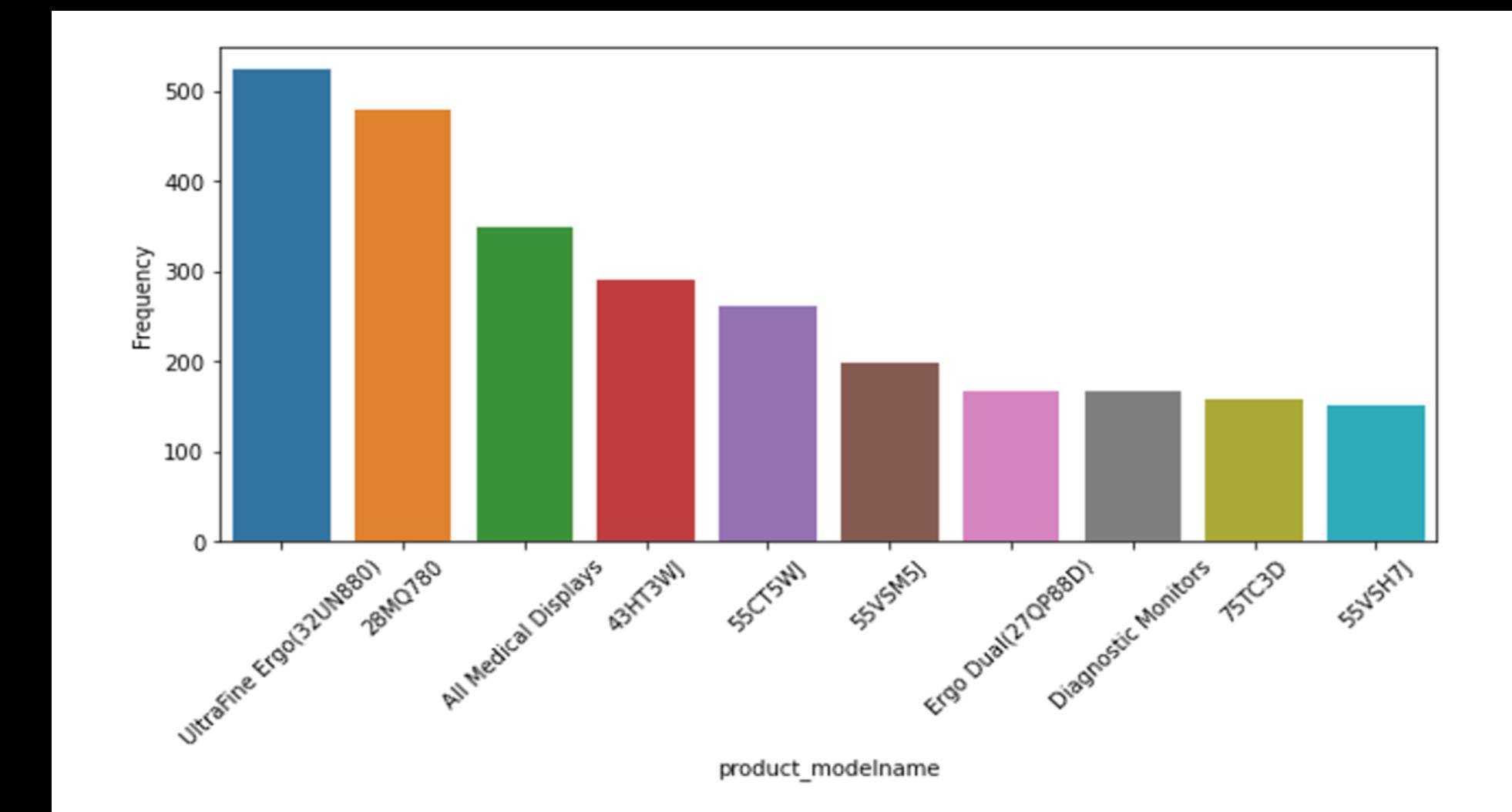


- 통계적 요약 - 범주형 변수

## product\_modelname 요청 제품 모델명

Info	Data
데이터 개수	9229개
결측치 개수	50070개 (84.44%)
고유값 개수	666개

[nan '98UH5J-H' '75XS4G' '86TR3DK' '43UR640S'  
 '86TR3DJ' '75UL3J-B' '22SM3G-B' '55VSM5J' '55UH5F-  
 H' '75TC3D' … '86TR3PJ' '65UR640S (ASIA)' '55VSH7J'  
 '43LT340C (EU)' '55LV77D' 'LAS009-F' '55UT640S  
 (ASIA)']

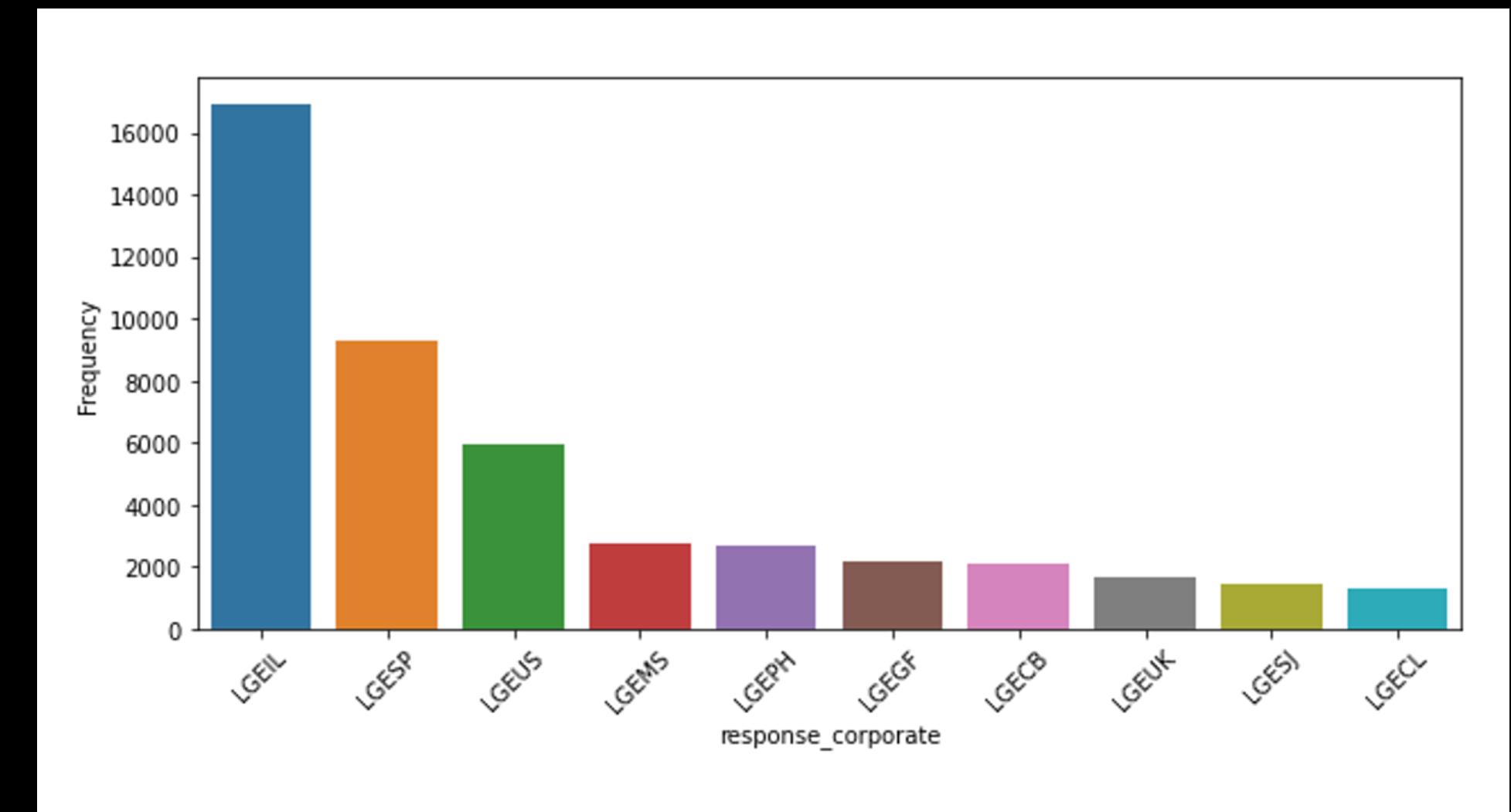


- 통계적 요약 - 범주형 변수

**response\_corporate**  
담당 자사 법인명

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	53개

[ 'LGEPH' 'LGEIL' 'LGEAF' 'LGESJ' 'LGESL'  
 'LGESP' … 'LGEGF' 'LGESA' 'LGEUS'  
 'LGECB' 'LGEMS' 'LGEEG' 'LGEEF' ]



- 통계적 요약 - 범주형 변수

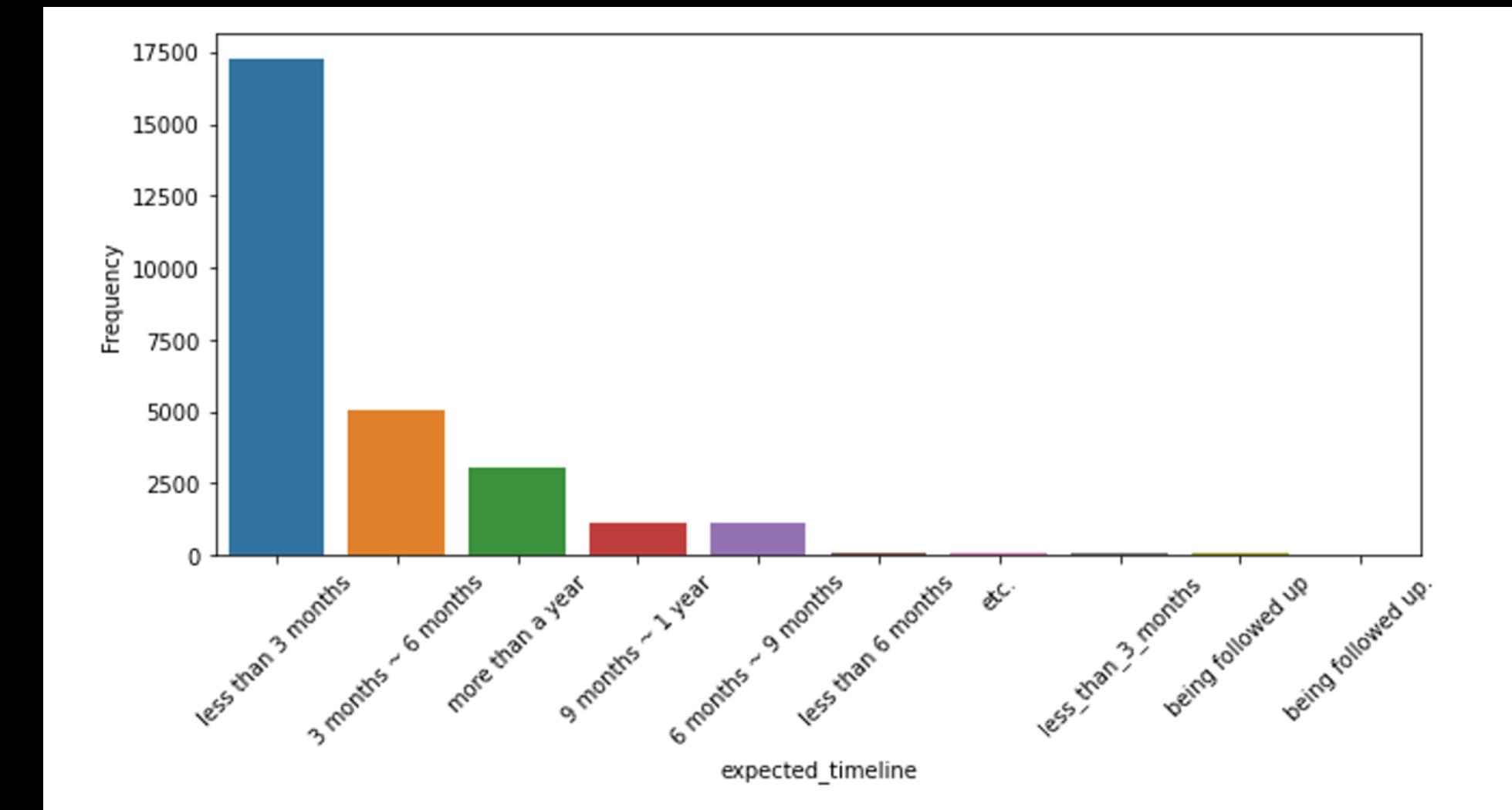
`expected_timeline`

고객의 요청한 처리 일정

대부분 특정 범주로 분리되어 있음, 이외의 구체적인 타임라인 데이터 존재

Info	Data
데이터 개수	28436개
결측치 개수	30863개 (52.02%)
고유값 개수	450개

[`'less than 3 months'` nan `'3 months ~ 6 months'` `'9 months ~ 1 year'` `'more than a year'` `'6 months ~ 9 months'` ... `'quote has been sent to customer.'` ]

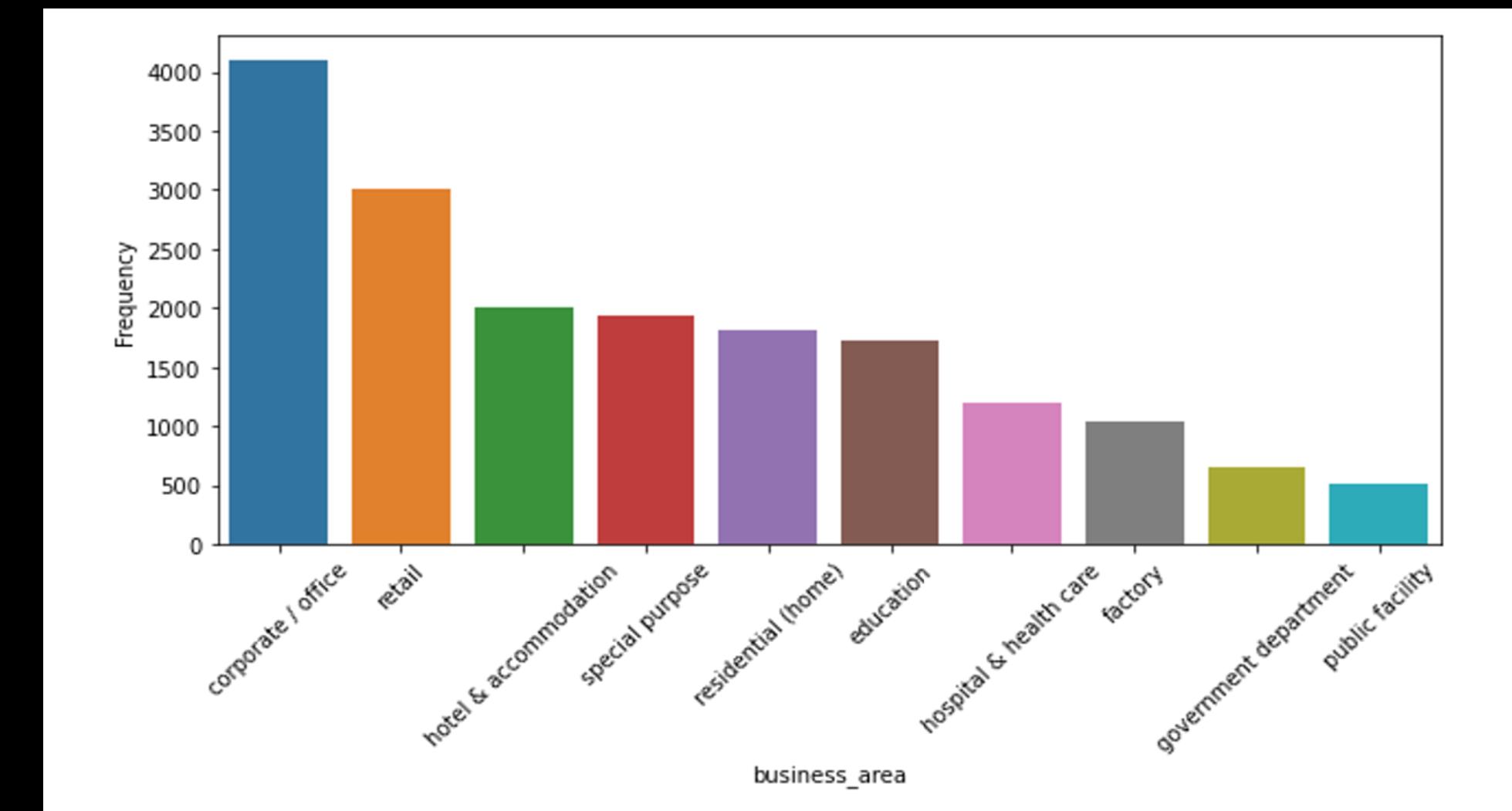


- 통계적 요약 - 범주형 변수

## business\_area 고객의 사업 영역

Info	Data
데이터 개수	18417개
결측치 개수	40882개 (68.92%)
고유값 개수	13개

['corporate / office' nan 'education' 'hotel & accommodation' ... 'government department' 'retail' 'factory' 'power plant / renewable energy' 'transportation' 'public facility']

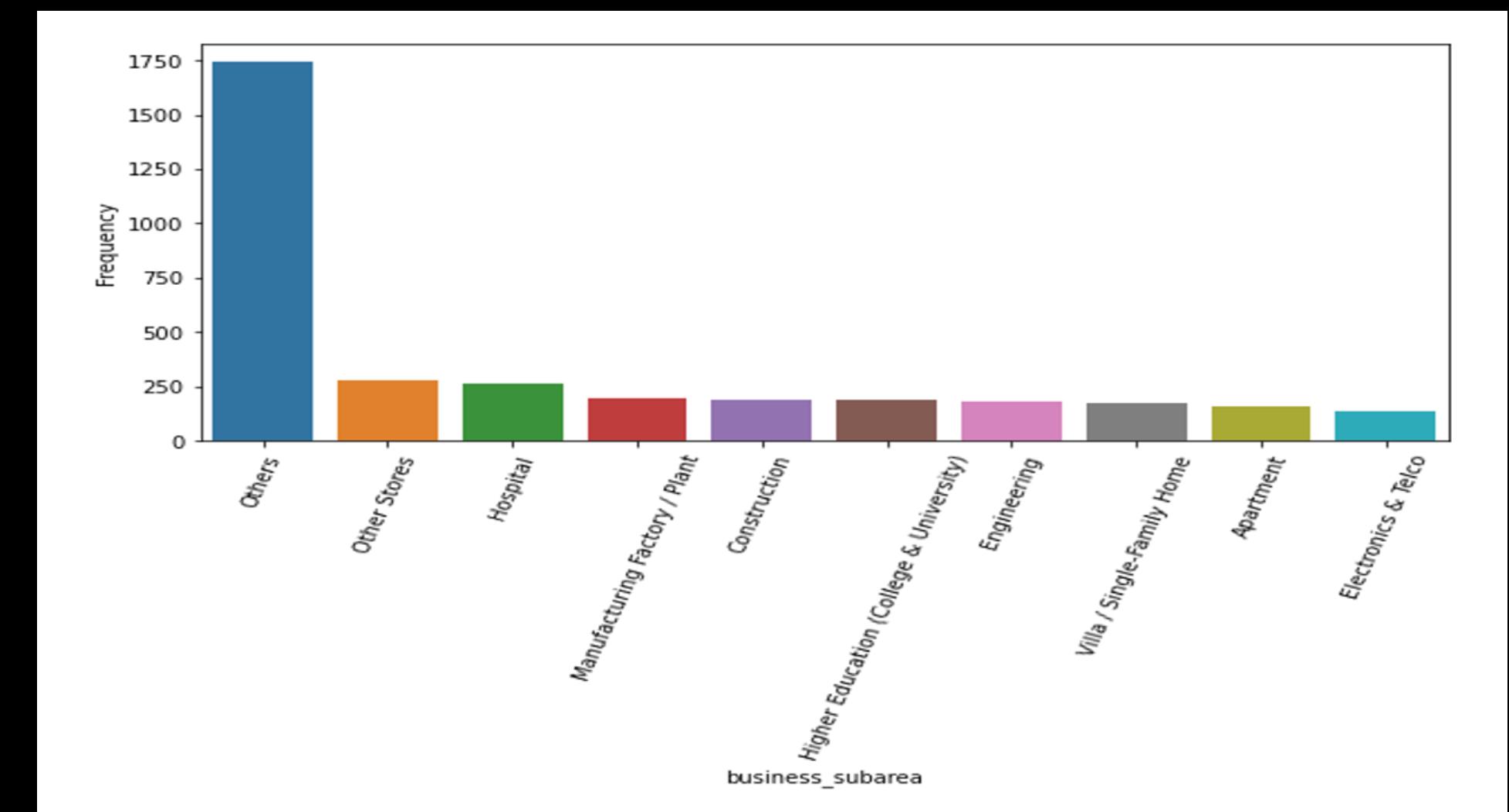


- 통계적 요약 - 범주형 변수

## business\_subarea 고객의 세부 사업 영역

Info	Data
데이터 개수	5526개
결측치 개수	53773개 (90.68%)
고유값 개수	87개

['Engineering' 'Advertising' 'Construction' 'IT/Software'  
 nan 'Manufacturing' 'Energy' 'Developer/Property' ...  
 'Entertainment' 'Agriculture' 'Pharmaceutical' 'Others'  
 'Banking' 'Consulting']



- 통계적 요약 - 범주형 변수

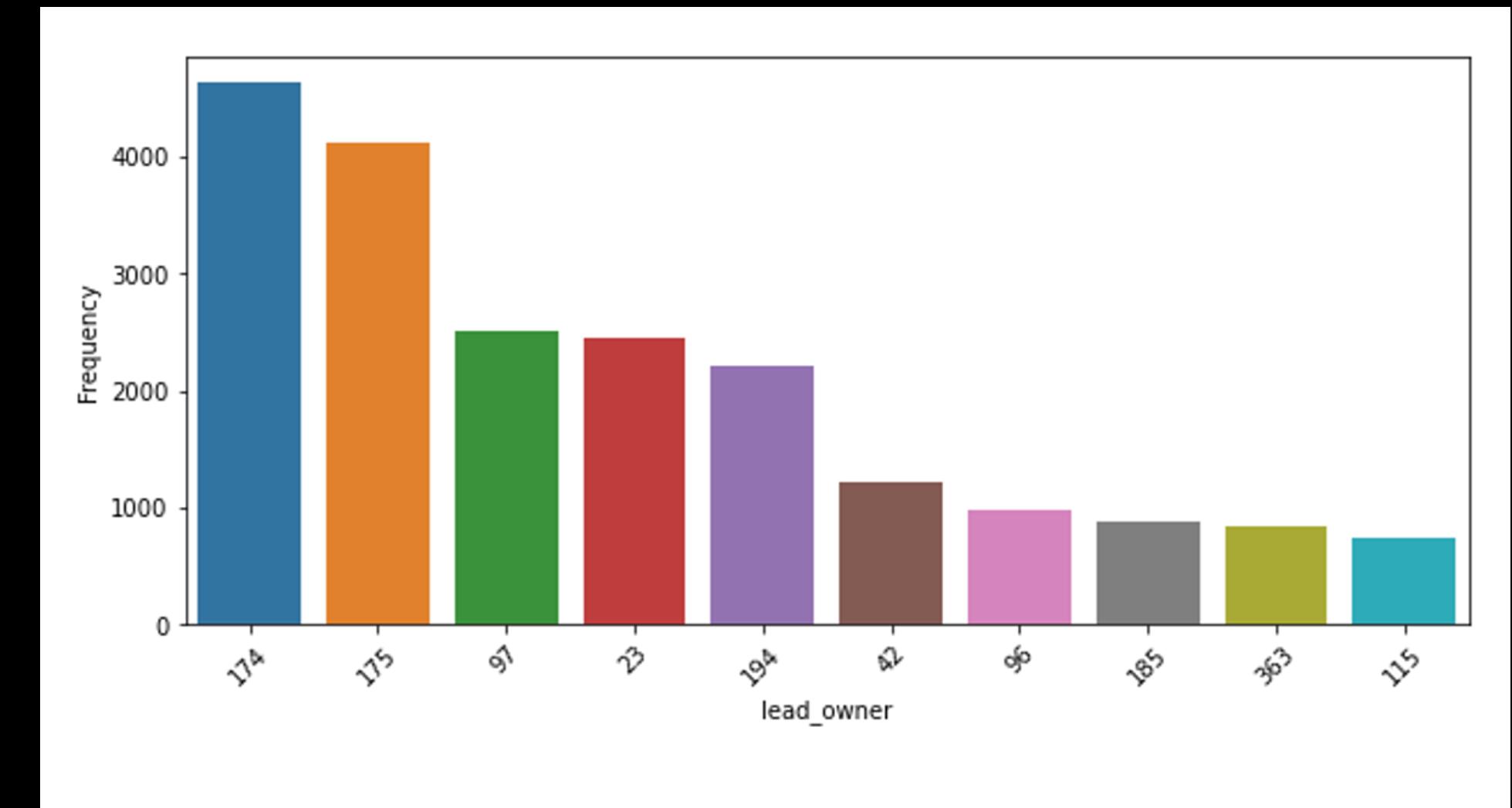
lead\_owner

영업 담당자 이름

타입이 int64이지만 의미에 따라 범주형으로 취급

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	984개

[ 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22  
23 24 25 ⋯ 775 1039 1099 1101 1102 1103 977 1105 1106  
1109 1110 964 1111 1114 ]



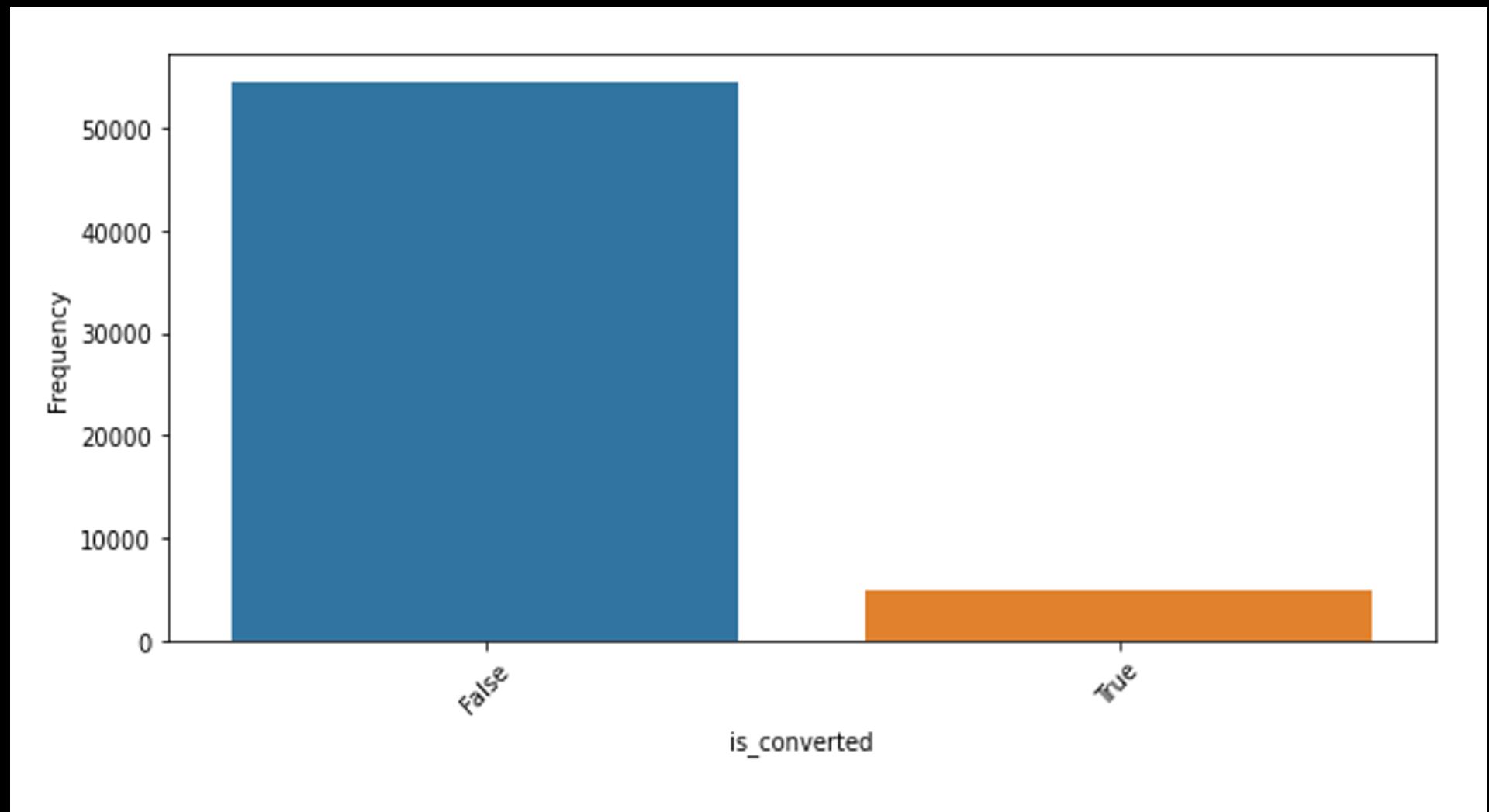
- 통계적 요약 - Target 변수

isConverted

영업 성공 여부. True일 시 성공  
데이터 불균형 존재

Info	Data
데이터 개수	59299개
결측치 개수	0개 (0%)
고유값 개수	2개

[ True False]



- 초기 통찰 및 발견

1. 가중치를 부여하는 변수들에서 높은 비율의 결측치 존재  
삭제/보완 필요
2. 범주형 변수들의 데이터 중 같은 의미이지만 다르게 표기된 값들이 존재  
표준화 작업 필요
3. ‘customer\_country’ 변수에서 슬래시(/) 기준으로 분리 가능  
새로운 설명 변수 도출 가능
4. ‘customer\_country’ 변수에서 국가명 표기법 다양함  
표준화 작업 필요
5. ‘expected\_timeline’ 변수에서 명확하게 구분되지 않은 timeline 존재  
키워드 추출을 통해 간소화된 카테고리로 재분류

- 결측치 처리

1. 결측치 처리

`id_strategic_ver, it_strategic_ver, idit_strategic_ver`

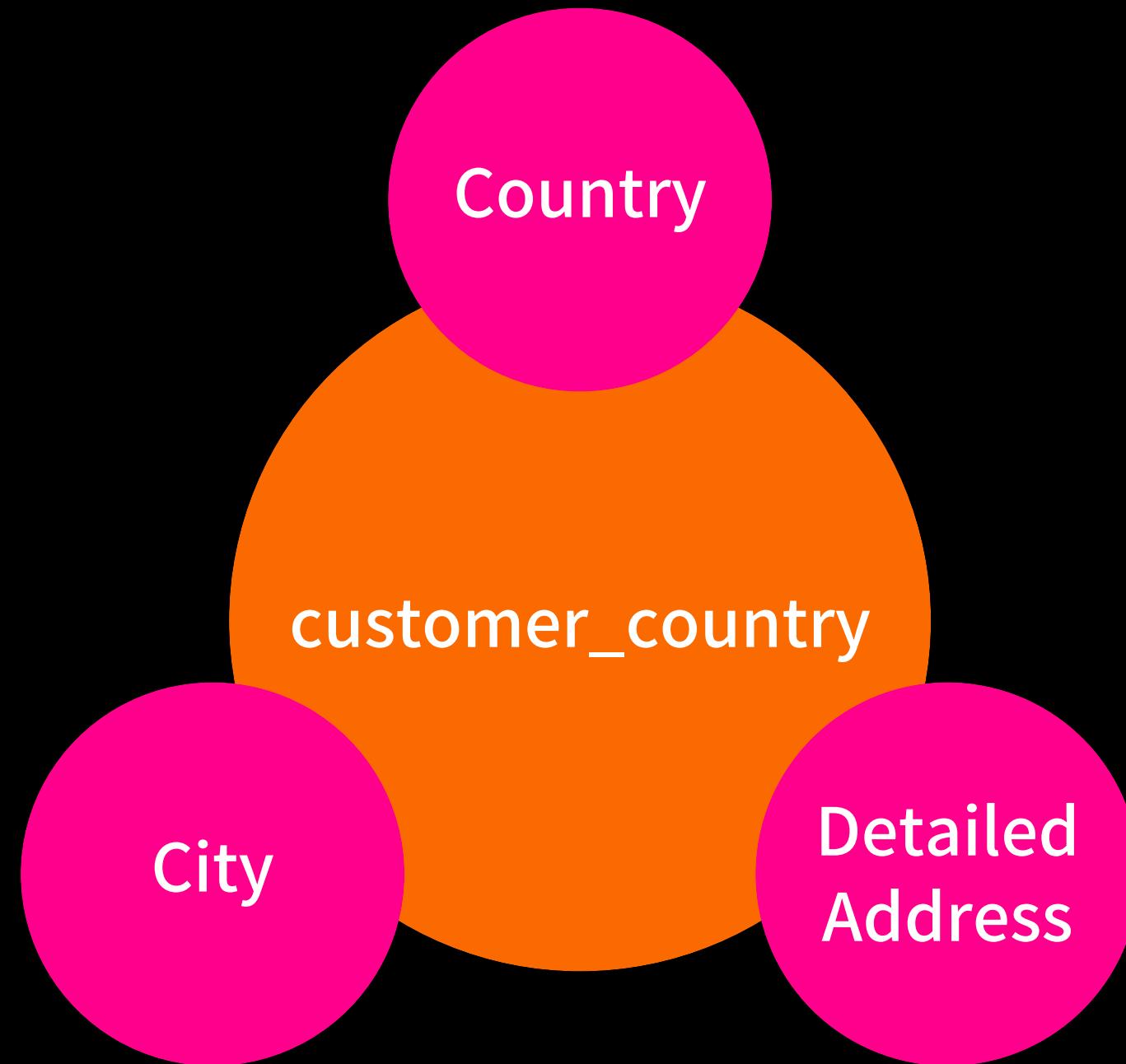
특정 사업부(Business Unit)에 대해 가중치를 부여해야 하지만 변수의 90% 이상이 결측치로  
존재 따라서 '**business\_unit**' == 'ID', 'IT'인 경우 1, 그렇지 않은 경우 0으로 결측치 대체

2. 결측치 제거

`business_subarea` (결측치: 90.68%)

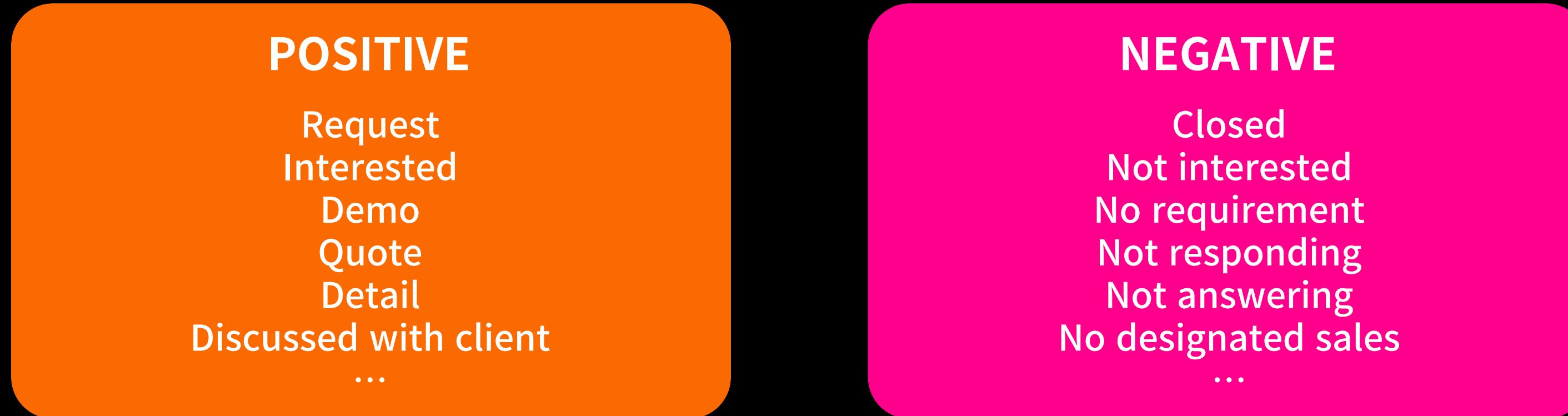
결측치 비율이 90% 이상인 변수 제거

- 데이터 표준화 - customer\_country



1. `customer_country` 변수를 슬래시(/) 기준으로 `country`, `city`, `detailed_address`로 분리하여 구체적인 설명 변수 추가 생성
2. 통일되어 있지 않은 `country` 변수의 국가명을 표준화  
Ex) "U.A.E": "United Arab Emirates"

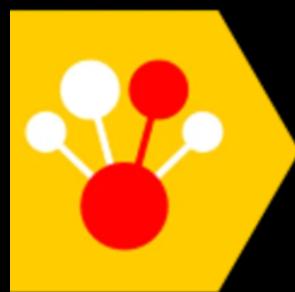
- 데이터 표준화 - expected\_timeline



특정 timeline으로 구분된 데이터 이외에 구체적으로 서술된 timeline 데이터 존재  
영업 전환이 성공적으로 이루어진 값(True)에서 positive 키워드를,  
영업 전환이 되지 않은 값(False)에서 negative 키워드를 추출  
각 키워드를 포함하고 있는 데이터를 ‘positive’ or ‘negative’로 대체

고유값의 수를 줄임으로써, 모델의 복잡성 감소하고 일반화 능력 향상

- CatBoostClassifier



# CatBoost

고차원 범주형 데이터의 효율적 처리

오버피팅 방지 및 안정적인 성능

하이퍼파라미터 튜닝

GPU 활용과 빠르고 효율적인 학습

여러 대표적인 분류 모델들(RandomForest, XGBoost, LightGBM 등)과의 성능을 평가 지표 (정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수 등)를 통해 평가하여 **CatBoostClassifier**를 최종 모델로 선정하였음

## ● Methodology

하이퍼파라미터	설명	영향
learning_rate	각 부스팅 단계에서의 단계 크기 조절	낮은 값: 학습 속도, 성능; 너무 낮으면 학습 부족 가능성
depth	결정 트리의 최대 깊이	깊이: 모델 복잡도, 오버피팅 위험
l2_leaf_reg	리프 노드의 가중치에 적용되는 L2 정규화 계수	오버피팅 방지
iterations	부스팅 단계의 수 또는 생성할 트리의 수	단계 수: 성능, 계산 시간, 오버피팅 가능성
border_count	수치형 피처 분할에 사용되는 경계의 수	경계 수 : 세밀한 분할 가능, 계산 시간
cat_features	범주형 피처의 인덱스	범주형 변수의 직접 처리로 성능 향상 가능
min_data_in_leaf	리프 노드가 되기 위해 필요한 최소 데이터 수	최소 데이터 수: 트리 깊이, 오버피팅 방지
bootstrap_type	부스팅 계산에 사용되는 샘플링 방법	데이터 샘플링 방식 차이를 통한 성능 최적화

팀 내 논의와 CatBoost 하이퍼파라미터에 대한 깊이 있는 이해를 바탕으로  
최적의 조합을 도출하기 위한 방법론을 구축하였음

- Hyperparameter Tuning

**GridSearchCV 기반 탐색**

- 가능한 모든 하이퍼파라미터 조합을 체계적으로 탐색
- 교차 검증을 통한 각 조합 성능 평가

**Confusion Matrix 기반 성능 분석**

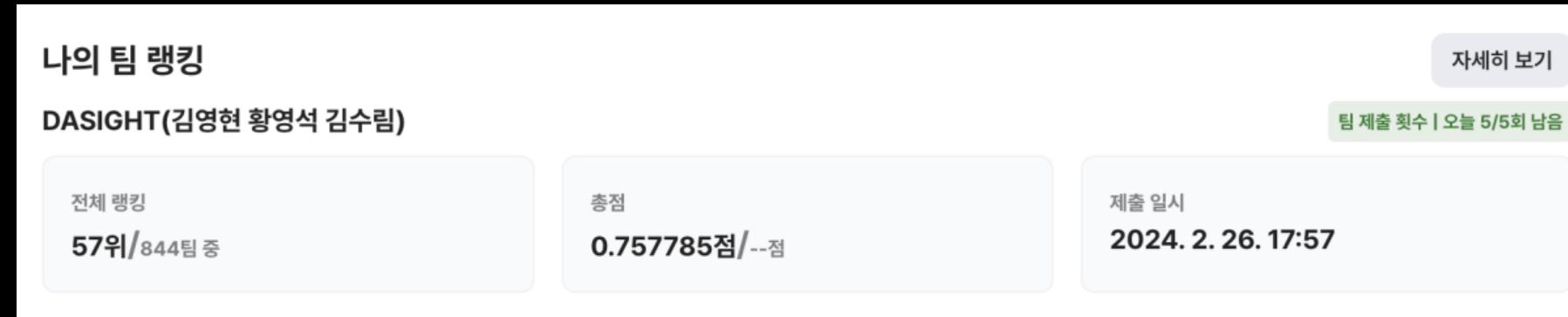
- 예측 결과 실제 결과를 비교
- 모델 성능 평가 및 특정 유형 오류 탐색

**Public Score 기반 검증**

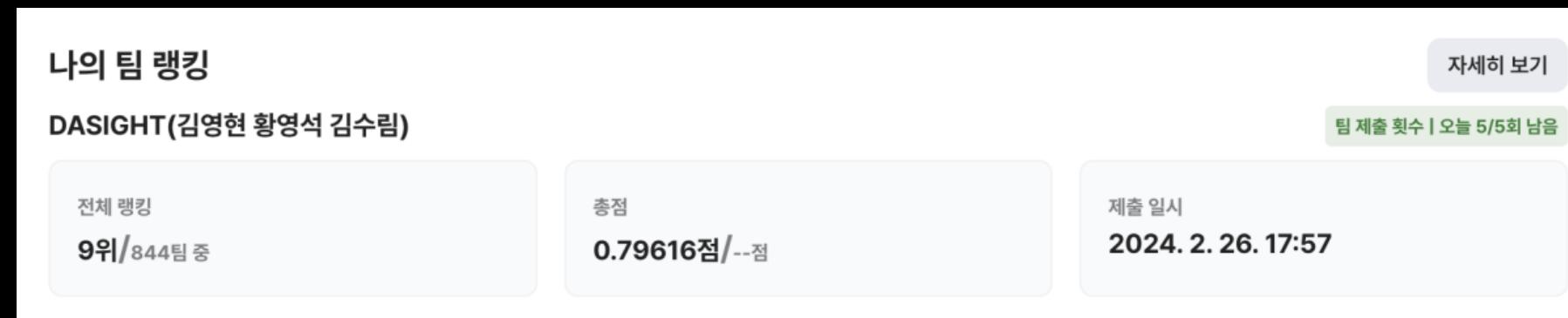
- 예측 성능 객관적 평가
- Feedback 기반 모델링 수정 및 개선

하이퍼 파라미터 튜닝은 GridSearchCV, Confusion Matrix, 및 Public Score를 활용하여 최적의 하이퍼파라미터 조합을 탐색하는 방법으로 수행하였음

## ● Conclusion



Public Score : 0.757785점  
Ranking : **57위**/884팀 중 (**상위 6.753%**)



Private Score : 0.79616점  
Ranking : **9위**/884팀 중 (**상위 1.066%**)

Public Score 대비 Private Score에서의 성능 향상이 두드러지게 나타나며  
모델이 미지의 데이터에 대해 뛰어난 일반화 능력을 보였음을 의미

## ● Improvements and Considerations

### 데이터 전처리 및 피처 엔지니어링

- 추가적인 탐색적 데이터 분석(EDA) 수행
- 피처 엔지니어링 강화

### 모델링 및 하이퍼파라미터 최적화

- 하이퍼파라미터 추가 탐색 및 재조정
- 고급 최적화 기법 활용

### 오버피팅 대비 전략

- 교차 검증 활용 강화
- 정규화 기법 추가 고려

성능 향상을 위한 추가적인 접근 방법을 고려함으로써  
더욱 개선된 모델 개발을 기대할 수 있음

감사합니다.