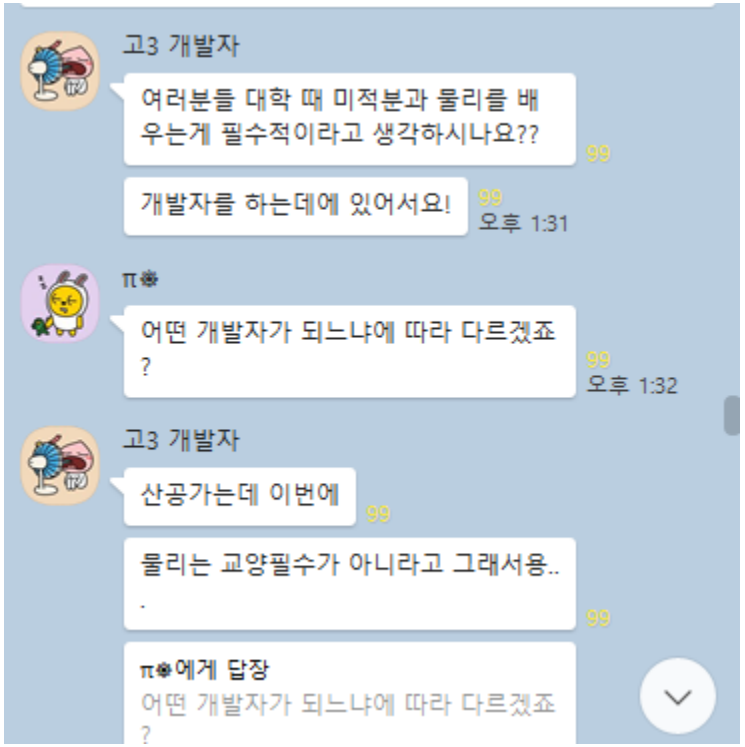
키워드, 시간대별 응답률 분석

BACKGROUND



무슨 주제로 이야기를 할까?

언제 이야기를 할까?

어떤 키워드로 질문을 하면 좋을까?

어떤 키워드를 내세워야 알맞은 답변을 받을 수 있을까?

BACKGROUND

GOAL

키워드, 시간대별로 응답률을 분석하여 정보를 더욱 효율적으로 얻는 전략 수립

DATASET

- [데이터리안] 데이터 분석 정보 공유방 (2023-08-13 ~ 2023-12-03)
  - 6,000 rows+

[데이터리안] 데이터 분석 정보 공유방 님과 카카오톡 대화
저장한 날짜 : 2023-12-03 16:42:51

--------------- 2023년 8월 13일 일요일 ---------------
ㅎ히님이 들어왔습니다.
빅지기/DA/2님이 나갔습니다.
--------------- 2023년 8월 14일 월요일 ---------------
[PA(어린이팬)/데이터분석/2년차] [오전 10:56] 군집별 CTR 데이터를 보고 있습니다.

특정 군집이 CTR이 가장 높은데, 해당 군집의 이벤트 수(노출)은 다른 군집에 비해 작은 편이에요.
'이 군집의 이벤트 수(노출)을 높히도록 유도하면 전체 CTR이 높이진다'라고 판단하기 위해서는 어떤 근거를 들어야할까요?
느낌 상, 집단의 수와 신뢰구간 등을 확인해봐야할 것 같은데, 명확히 어떤 방법을 통해 확인해야할지 감이 안와서 질문드립니다!
큠/데이터분석님이 나갔습니다.
[어린이/데이터분석/10년차] [오후 4:01] 저라면 이벤트 수는 적지만 다른 군집보다 CTR이 유의미하게 차이가 있는지 먼저 검증하고,

해당 군집과 유사한 군집들은 이런이런 애들이 있으니 노출해서 CTR 증대가 가능할 것 같다고 정리할 것 같네요
[어린이/데이터분석/10년차] [오후 4:03] 제가 이해한게 맞나요??
[어린이/데이터분석/10년차] [오후 4:04] 대다수 근데 원래 숫자이 작으면 컨버전 과정에서 비율 바이어스가 엄청껴요
[어린이/데이터분석/10년차] [오후 4:04] 실무적으로 진행해보시면 아시겠지만요..
[어린이/데이터분석/10년차] [오후 4:04] 이-> 가
업무하면서쳐서 오타많네요 이해해주세요
[어린이/데이터분석/10년차] [오후 4:05] 노출한다는게 특히 광고쪽이면 더더욱. 집단의 규모 자체가 작을 수도 있으니 유념해서 보셔야합니다
[Roy/PM/5년차] [오후 4:08] 각 군집별 모수가 어떻게 됐나요? CTR 높았던 군집의 모수는 어느정도인지도 궁금합니다
[어린이/데이터분석/10년차] [오후 4:09] 몇천단위 정도 되려나요? ㅎㅎ 아무래도 반응율도 그렇고요.

PIPE LINE

Data Preprocessing → EDA → Application → Conclusion

DATA PREPROCESSING

DATA PREPROCESSING

Raw Data

입장/퇴장

닉네임

지미/데이터분석/5년차님이 들어왔습니다.
라따구리/경영학전공님이 나갔습니다.
레비/그로스마케터/2년차님이 들어왔습니다.
Jess/CRM/2년차님이 들어왔습니다.
[파이/Growth/0년차] [오후 6:22] 안녕하세요! 고객 평생가치를 계산할 때 주로 업계에서 사용하는 공식이 뭔지 궁금합니다...!
또, 개인 프로젝트로 LTV를 고객 세그먼트별로 분석해보려고 하는데 ARPPU가 지표로 사용될 때, LTV를 계산해서 얻는 효용이 얼마나 의미있는지도 궁금합니다.
아임비타/CRM/10년차님이 들어왔습니다.
[라이언/기계] [오후 6:35] prods 딸만한가요? 혹시 해보신분 있을까요? sqld나 adsp, 빅분기에 비해 자료가 많지 않네용
하고파/서비스기획/5년차님이 들어왔습니다.
[블루냥/데이터사이언스] [오후 6:37] 삭제된 메시지입니다.
[블루냥/데이터사이언스] [오후 6:38] 죄송합니다 잘못 눌렀네요;
도마/그로스/8년차님이 들어왔습니다.
보리/마케팅/3년차님이 들어왔습니다.
코오카/마케팅/5년차님이 들어왔습니다.
캉캉/CRM/6년차님이 들어왔습니다.
숭숭/마케팅/4년차님이 들어왔습니다.

시간

[클릭/데이터분석/1년차] [오후 8:11] 1. 모델링 할거 아니면 arppu/이탈율을 일/주/월간으로요
2. 객단가가 ltv를 넘어서는 안되겠죠. 세그먼트를 광고 채널로 맞추고 roas랑 비슷하게 씁니다
[베개를 부비적대는 라이언/통계] [오후 8:14] 안녕하세요. 포트폴리오 관련해서 질문드립니다. 혹시 포폴 제출할 때 회사가 집중적으로 보는 능력만 추려서 포폴을 제출하시나.

제가 한 경험들이 시각화+머신러닝+딥러닝인데, 회사에서 원하는 일이 머신러닝 쪽에 집중되어 있으면 제출 포폴에서 시각화 및 딥러닝 부분은 제외하는게 맞는건지 제가 했
[스티브/데이터분석/2년차] [오후 8:51] 서류 평가하는 사람마다 다르겠지만 저는 이 사람이 JD와 얼마나 fit한지 우선적으로 보기 때문에 최대한 머신러닝 포폴을 강조한 사람
혹시나 서류 통과하시고 인터뷰할 기회가 있으시면 시각화나 딥러닝 포폴을 하드카피로 제출해보는걸 추천 드려봅니다
졸린 무지님이 나갔습니다.
[베개를 부비적대는 라이언/통계] [오후 9:15] 너무 고민되었는데 답변 정말 감사합니다!
피치/데이터님이 들어왔습니다.
[파이/Growth/0년차] [오후 11:12] 감사합니다!!
[Frodo/데이터분석/신입] [오후 11:45] 안녕하세요! 혹시 보통 공백기는 어느 정도까지가 그나마 괜찮은 정도라고 볼 수 있을까요..??

채팅 내용

DATA PREPROCESSING

1 데이터 정형화

	talk_date	day_name	writer	wrote_at	msg	action_msg	is_talking_activity	is_notice_action	is_deleted_msg	is_emoji	is_picture	is_search
0	2023-08-13	일요일	ㅎ히	NaN	ㅎ히님이 들어왔습니다.	들어오기	False	False	False	False	False	False
1	2023-08-13	일요일	빅지기/DA/2	NaN	빅지기/DA/2님이 나갔습니다.	나가기	False	False	False	False	False	False
2	2023-08-14	월요일	PA(어린이팬)/데이터분석/2년차	10:56	군집별 CTR 데이터를 보고 있습니다. 특정 군집이 CTR이 가장 높은데, 해당 군...	NaN	True	False	False	False	False	False
3	2023-08-14	월요일	금/데이터분석	NaN	금/데이터분석님이 나갔습니다.	나가기	False	False	False	False	False	False
4	2023-08-14	월요일	어린이/데이터분석/10년차	16:01	저라면 이벤트 수는 적지만 다른 군집보다 CTR이 유의미하게 차이가 있는지 먼저 검...	NaN	True	False	False	False	False	False

Variable

Name		Name	
talk_date	작성날짜(yyyy-mm-dd)	is_talking_activity	채팅여부
day_name	요일	is_deleted_msg	채팅삭제여부
writer	작성자	is_emoji	이모티콘 사용여부
wrote_at	작성시간	is_picture	그림(사진) 사용여부
msg	채팅내용	is_search	검색(#) 여부
action_msg	• 들어오기, 나가기, 메시지 가리기 3개의 항목으로 구성 • 들어온 인원의 경우 0, 나간 인원의 경우 1, 메시지 가리기인 경우 2로 분류 진행		

2 키워드 분석을 위한 Hand-Labeling

	writer	wrote_at	msg	group_num	msg_label
0	PA(어린이팬)	10:56:00	군집별 CTR 데이터를 보고 있습니다. 특정 군집이 CTR이 가장 높은데, 해당 군...	1.0	q
1	어린이	16:01:00	저라면 이벤트 수는 적지만 다른 군집보다 CTR이 유의미하게 차이가 있는지 먼저 검...	1.0	a
2	Roy	16:08:00	각 군집별 모수가 어떻게 됐나요? CTR 높았던 군집의 모수는 어느정도인지도 궁금합니다	1.0	a
3	어린이	16:09:00	몇천단위 정도 되려나요? ㅎㅎ 아무래도 반응율도 그렇고요. 사실 광고면 이 타겟 확...	1.0	a
4	PA(어린이팬)	16:15:00	답변 감사합니다! 이렇게 이해했는데 맞을까요? 1. 전체 집단 vs (CTR이 높았...	1.0	c

- **group_num**
  - 특정 질문에 대한 대화를 그룹핑
  - 다른 주제에 대해서 질의응답을 한 경우 다른 group_num이 부여

- **msg_lab**

Variable	name	
q	question	질문자의 최초 질문
a	answer	답변
c	check	질문자의 확인(재질문, 감사합니다 등 포함)

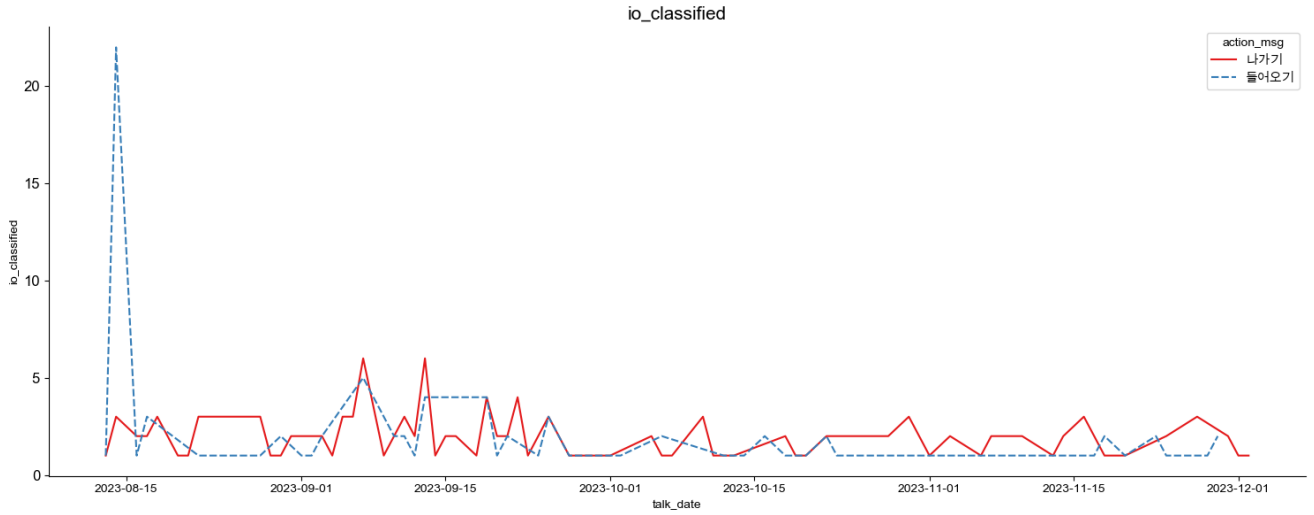
Exploratory Data Analysis

Exploratory Data Analysis

1 유입/이탈 인원 파악

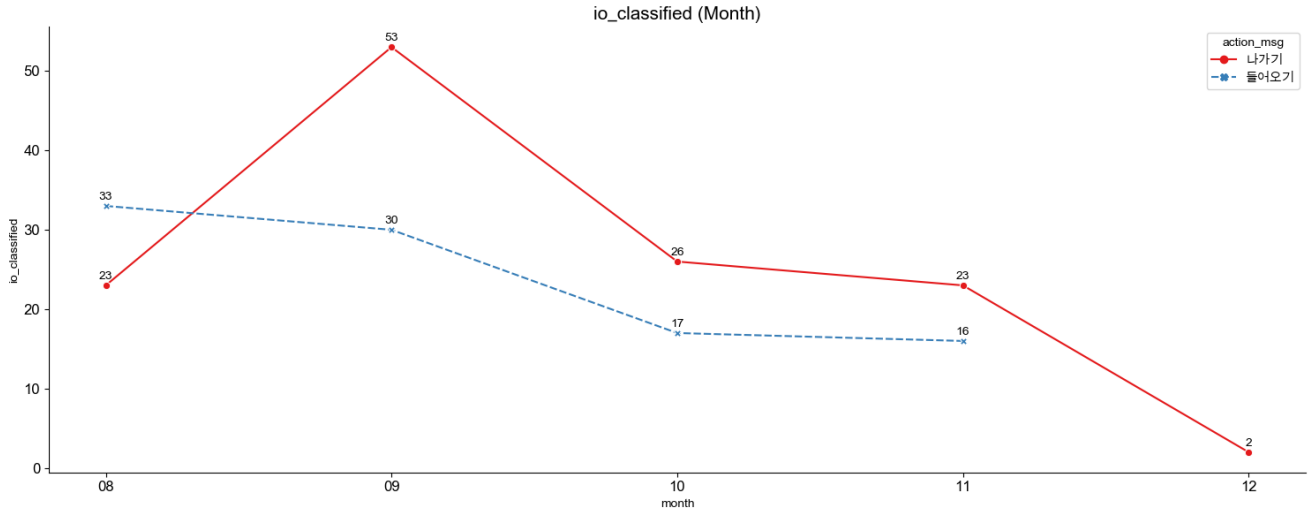
1-1. 월별 전체 활동량 파악

- Red solid line : Out
- Blue dotted line : In



1-2. 월 단위 유입/이탈 인원 파악

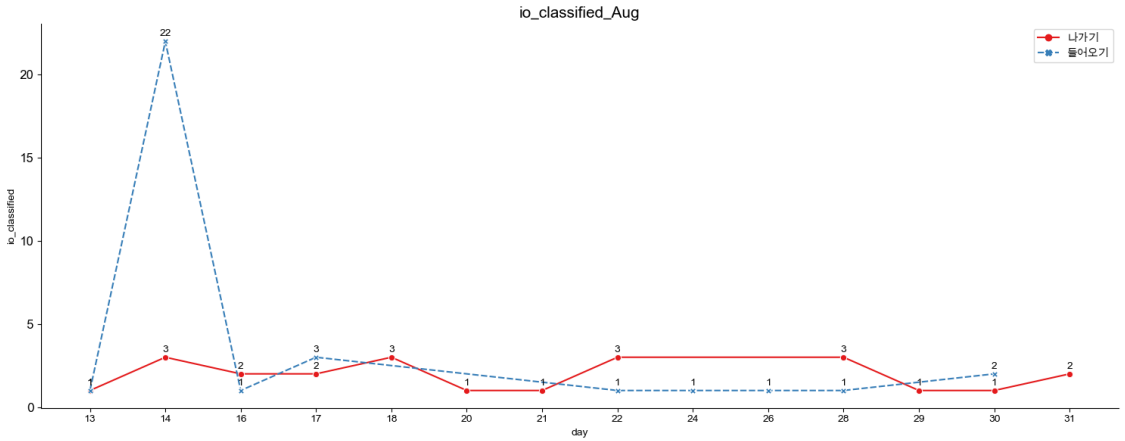
- 월(month)을 기준으로 전체 유입/이탈 인원 count
- 채팅방에 들어오는 인원은 **지속적으로 감소하는 추세를 보임**
- 12월 이상적으로 낮은 count를 확인

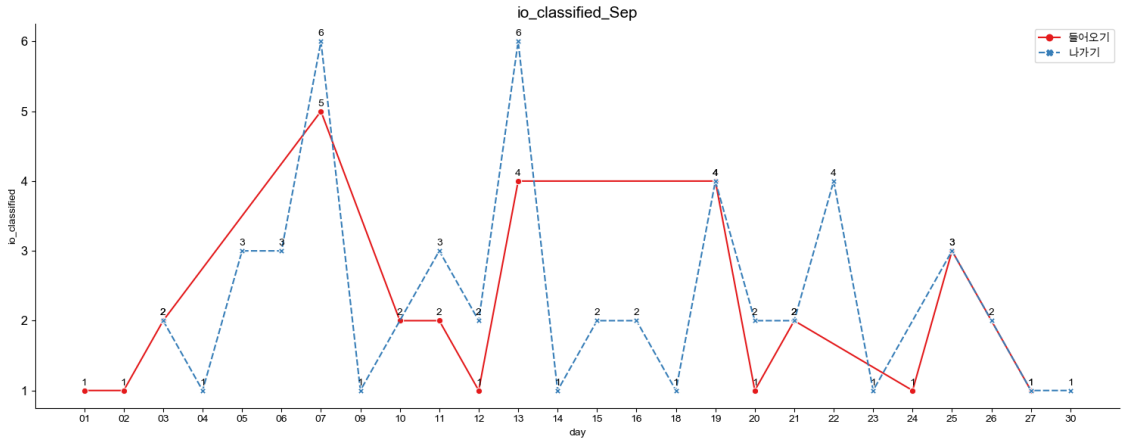


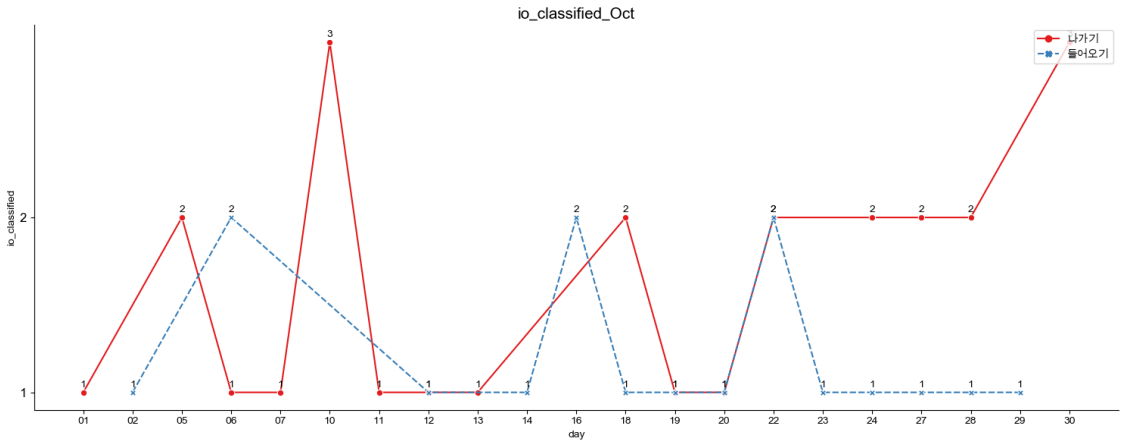
Exploratory Data Analysis

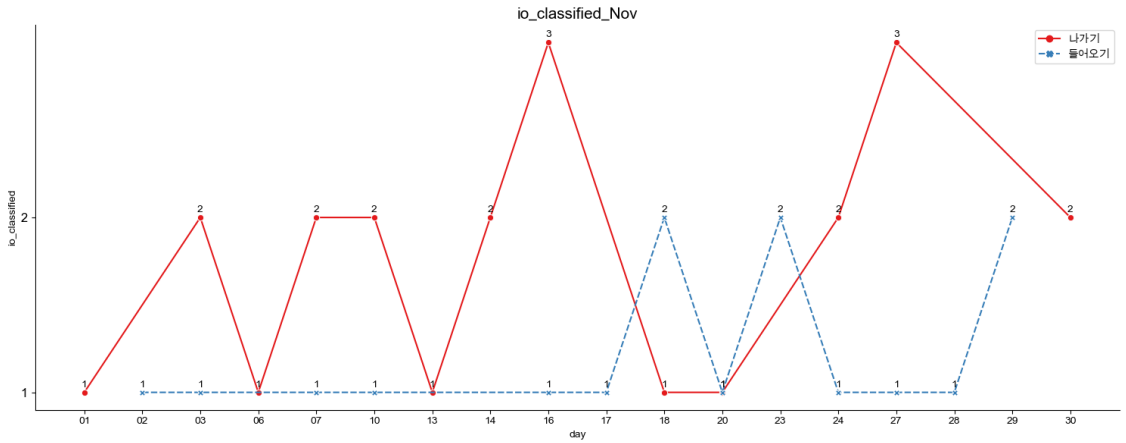
1 유입/이탈 인원 파악

1-3. 월(Month)에 따른 일(Day)별 유입/이탈 인원 파악





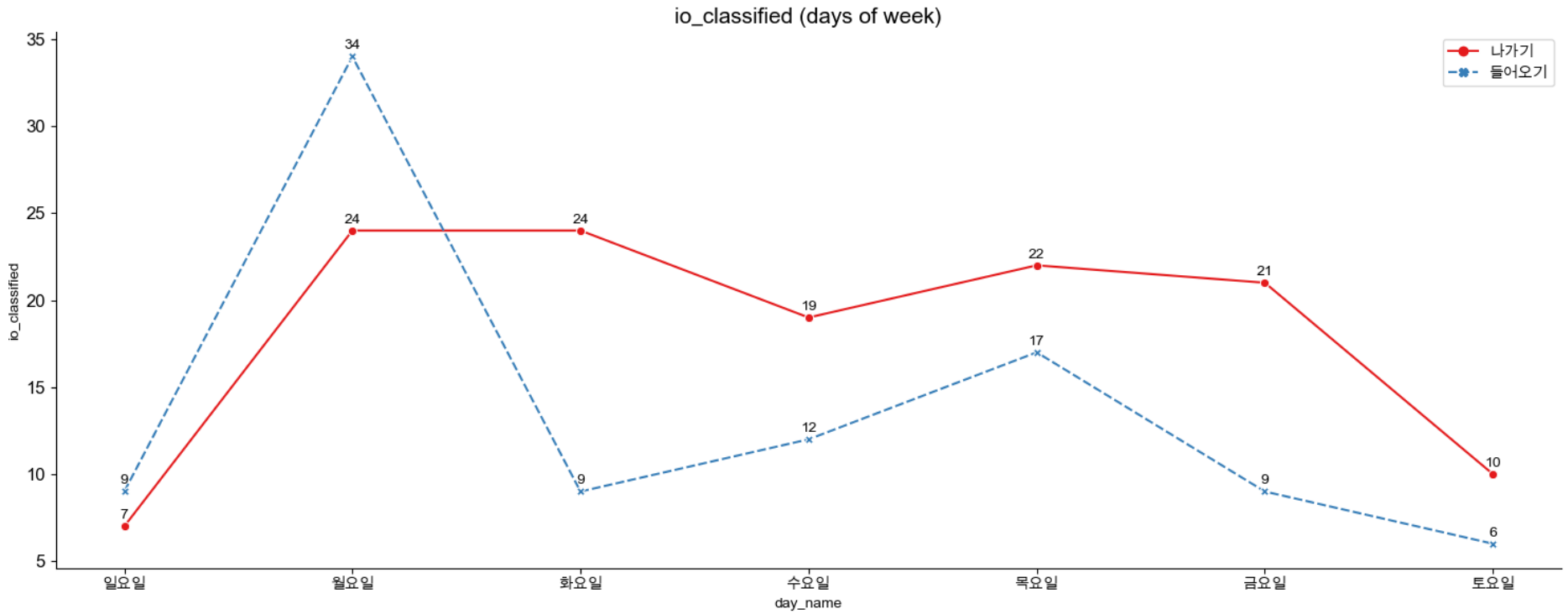




1 유입/이탈 인원 파악

1-4. 요일별 유입/이탈 인원 파악

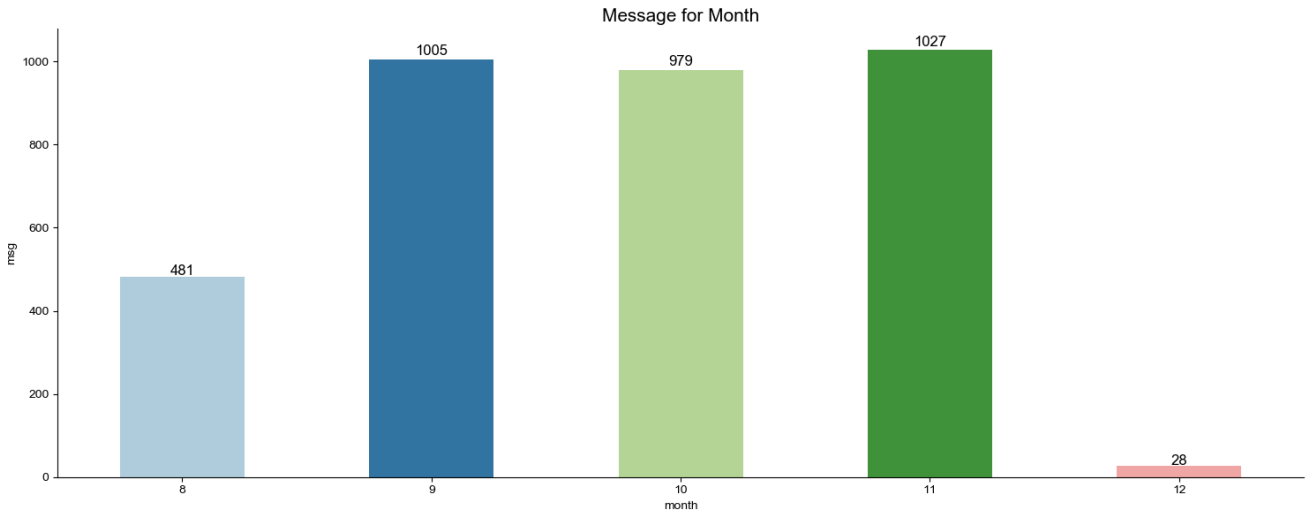
- 평균적으로 **월요일**에 유입/이탈이 활발한 것을 확인
- 주말(토, 일)요일의 경우 **유입/이탈 인원이 확연히 감소**



2 메세지 갯수를 통한 활동량 파악

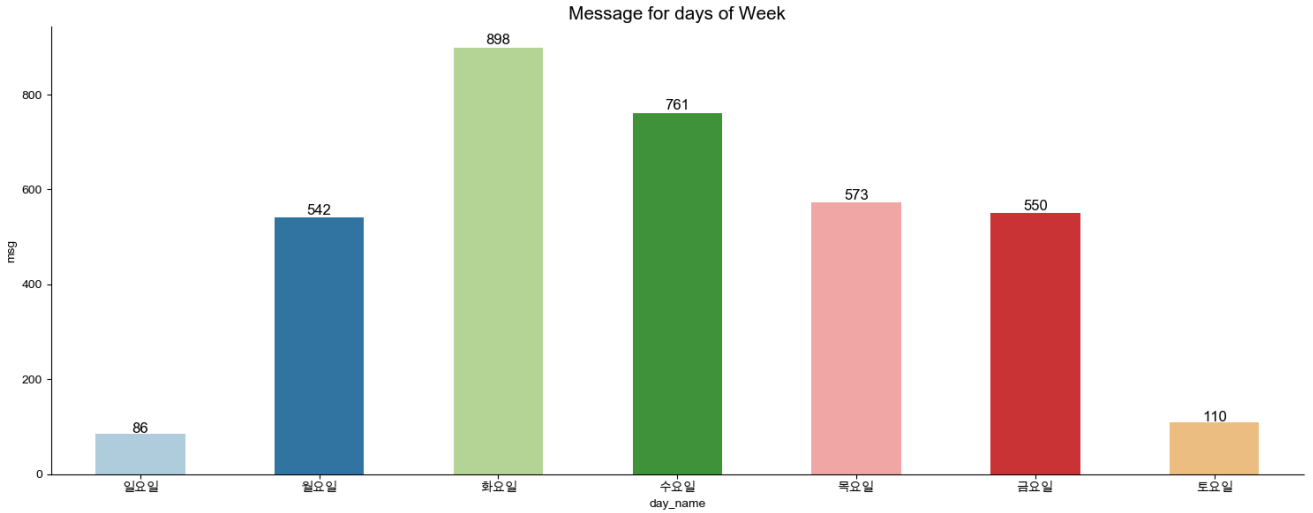
2-1. 월별 전체 활동량 파악

- 8월, 12월의 경우 기간의 차이로 인해 적게 나타남
- 전체적으로 비슷한 활동량을 보임



2-2. 요일별 전체 활동량 파악

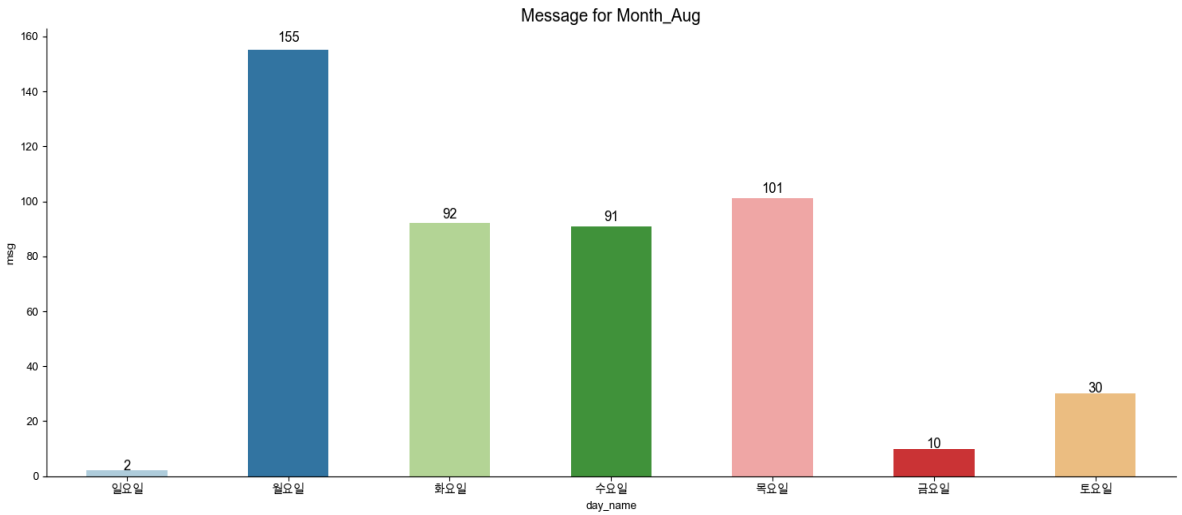
- 주중(월, 화, 수, 목, 금)의 경우 활동이 활발하게 일어남
- 주말(토, 일)의 경우 활동량이 미비함

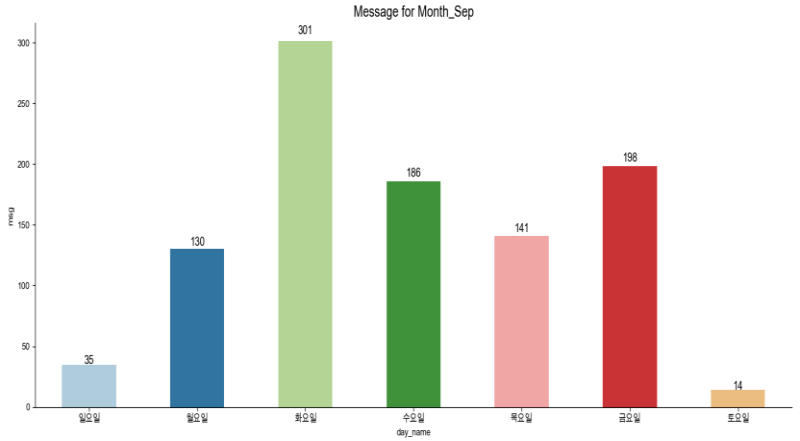


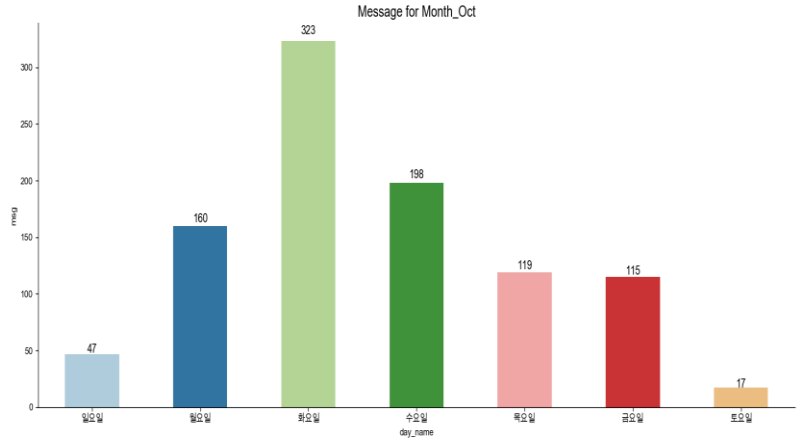
2 메세지 갯수를 통한 활동량 파악

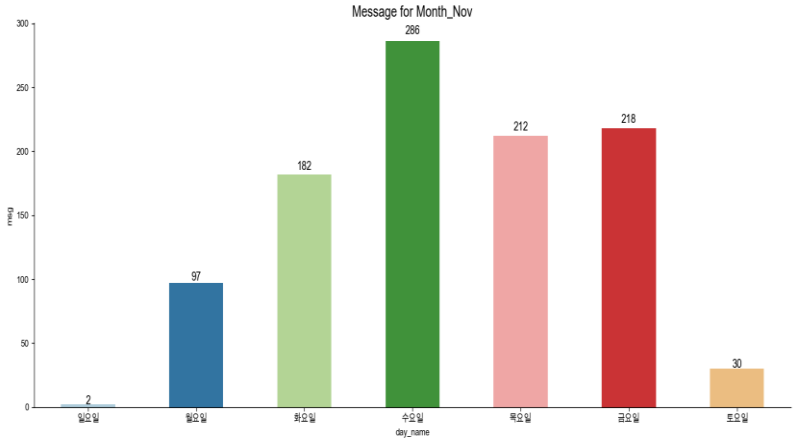
2-3. 월별 요일별 활동량 파악

- 전체적으로 주중에 공통적으로 활발하게 활동하는 것으로 나타남
- 그러나, 특정 요일에 활발하다는 결론은 내지 못함





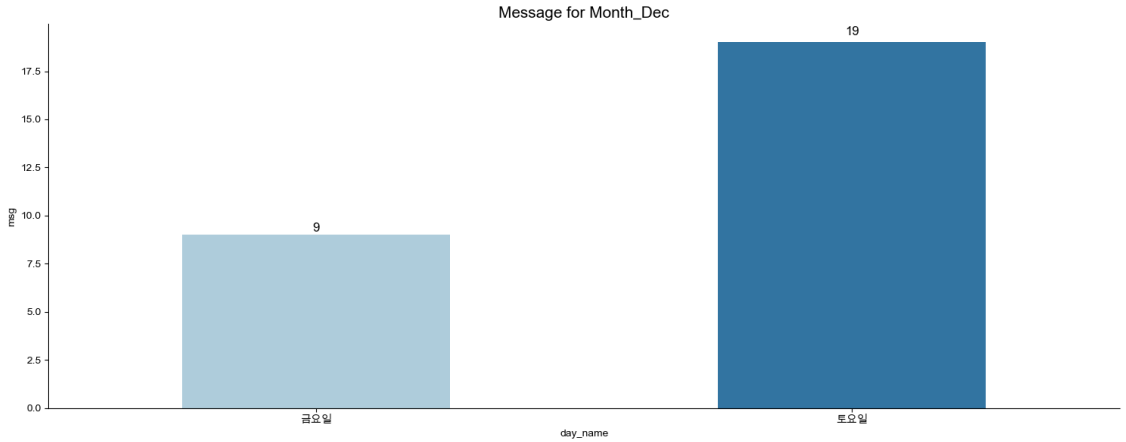




Exploratory Data Analysis

2 CASE : December

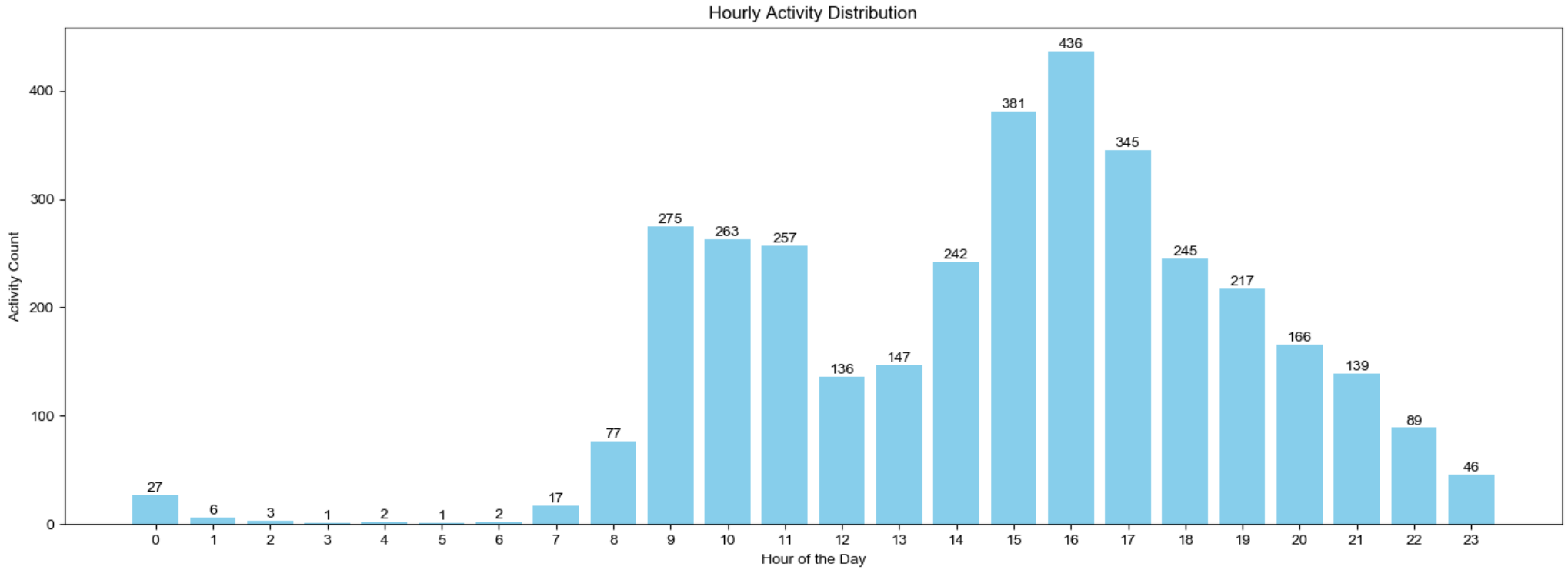
- 12월의 경우 데이터가 23-12-02까지 존재하여 count 수가 낮음



2 메세지 갯수를 통한 활동량 파악

2-4. 시간대별 활동량

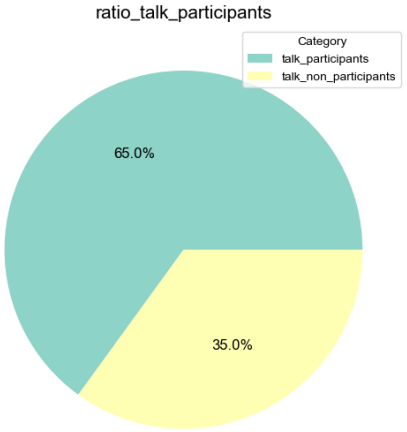
- 활동시간(9시~18시) 내 커뮤니케이션에 활발하게 일어남
- 특히, 15~17시 가장 높은 분포를 보임

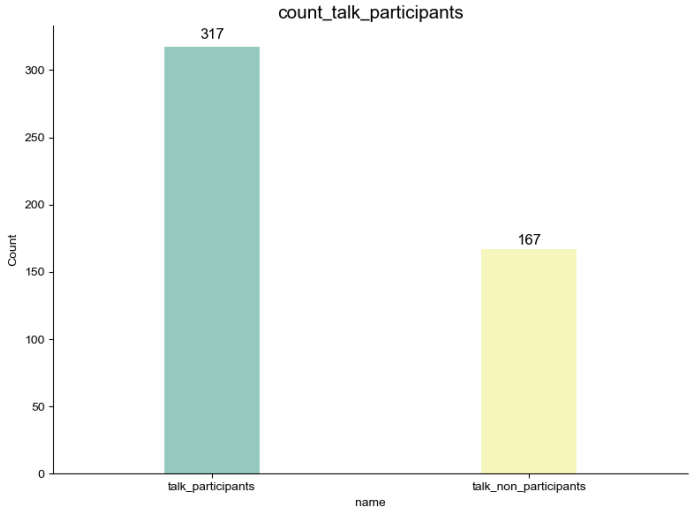


3 채팅 참여자 파악

3-1. 채팅 참여 여부 판단

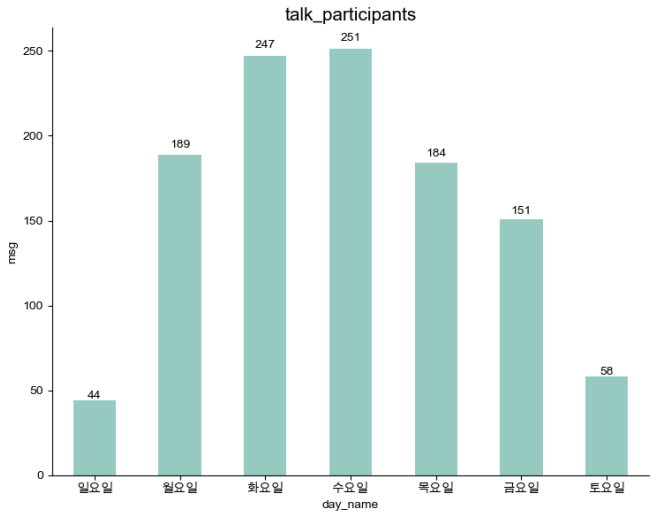
- 채팅방 내 채팅에 참여한 사람과 참여하지 않은 사람의 비율을 파악
- 참여자의 경우 317명(65%), 미참여자의 경우 167명(35%)으로 나타남

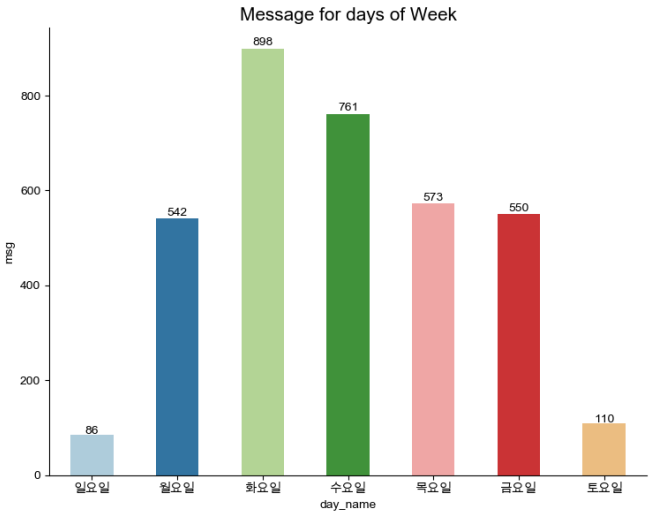




3-2. 요일별 채팅 참여 여부 파악

- 요일별 채팅 참여자 수 파악
- 채팅 참여의 경우 주중에서도 화요일, 수요일에 특히 활동이 많은 것을 확인





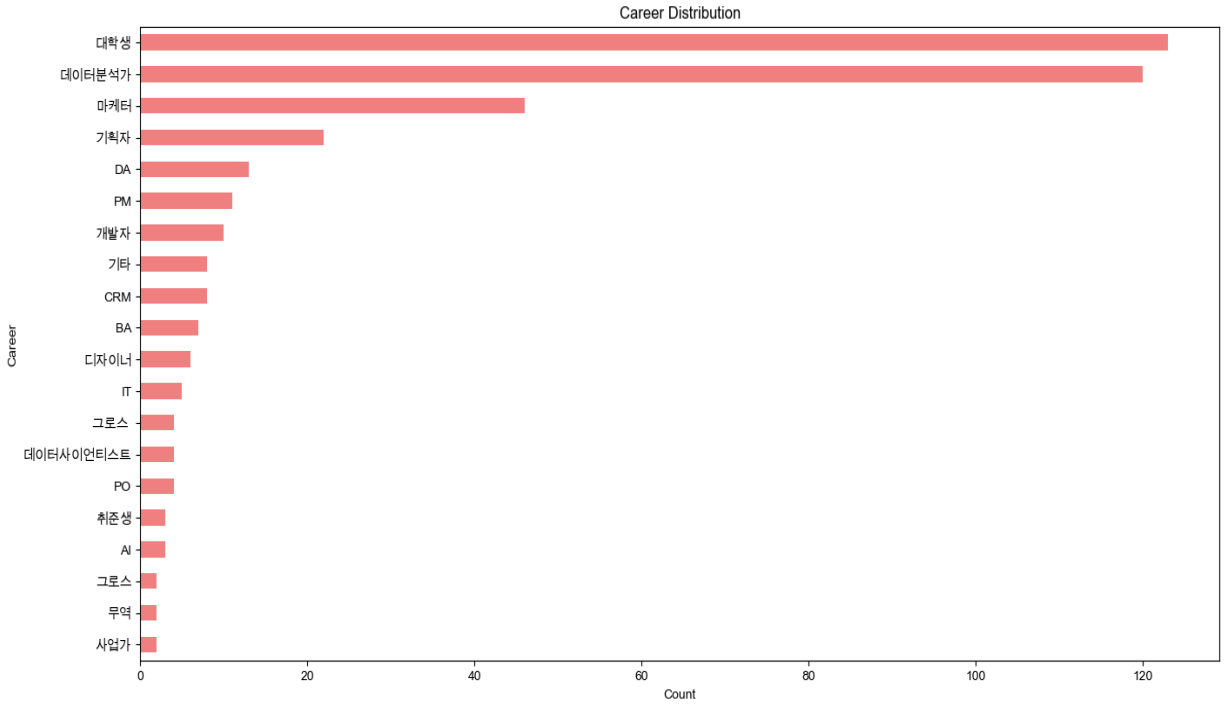
3 채팅 참여자 파악

3-3. 직군별 참여자 파악

- 직군 분류 2차 가공 진행

	writer	jungong	
0	ㅎ히		
1	빅지기	DA	DA
2	PA(어린이팬)	데이터분석	데이터분석가
3	귬	데이터분석	데이터분석가
4	어린이	데이터분석	데이터분석가
5	Roy	PM	PM
11	도달	DA	DA
22	나나	crm	CRM
23	마케터	4년차	마케터
24	Arrogant Jay-G		
25	푸디	데이터분석	데이터분석가
26	리미	CRM	CRM
27	무지무지	crm마케터	CRM
28	에르난데스		
29	애드	crm데이터분석	CRM
30	데이터초보	마케터	마케터
31	냠냠		
32	부쓰	데이터분석	데이터분석가
33	그로쓰	그로스	그로스
34	건배	마케팅	마케터
35	지미	데이터분석	데이터분석가

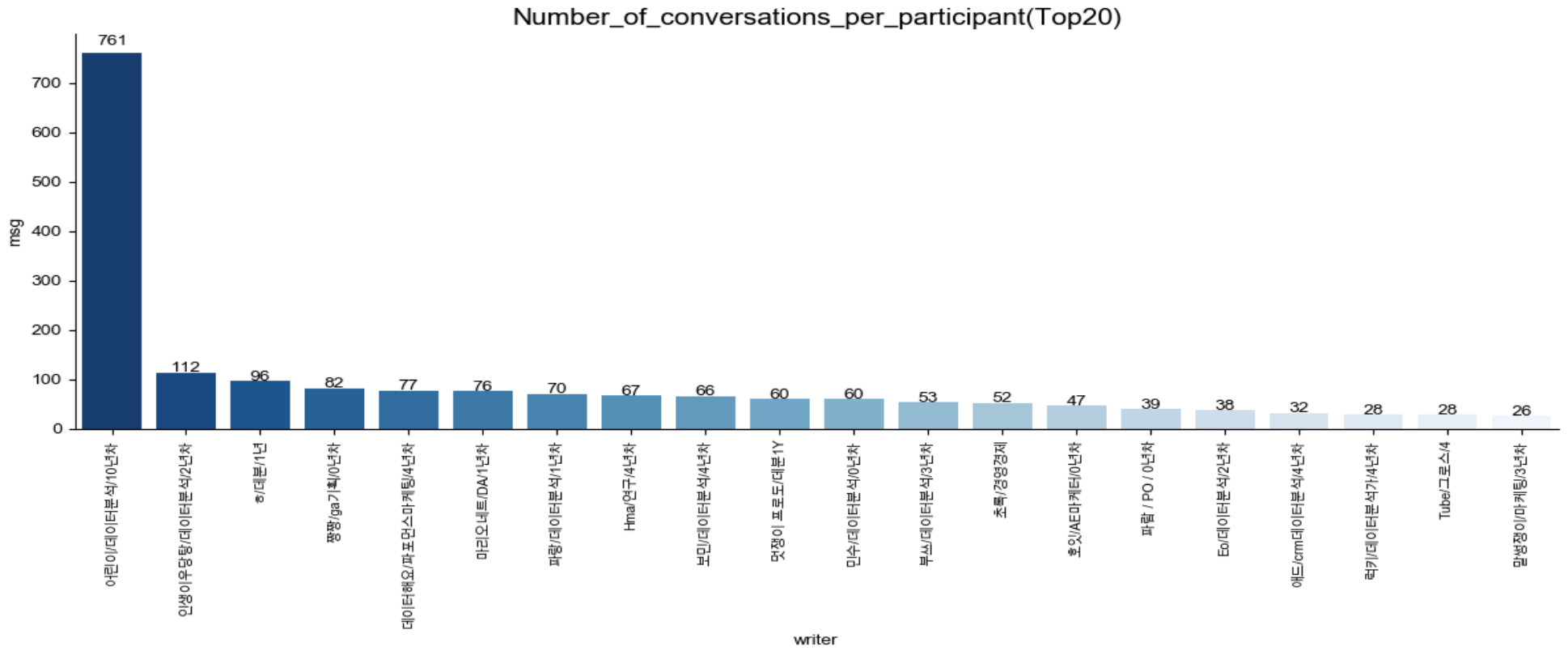
- 대학생, 데이터분석가, 마케터 직군 순으로 많은 참여자가 분포



3 채팅 참여자 파악

3-4. 참여자 별 대화내용 수 파악

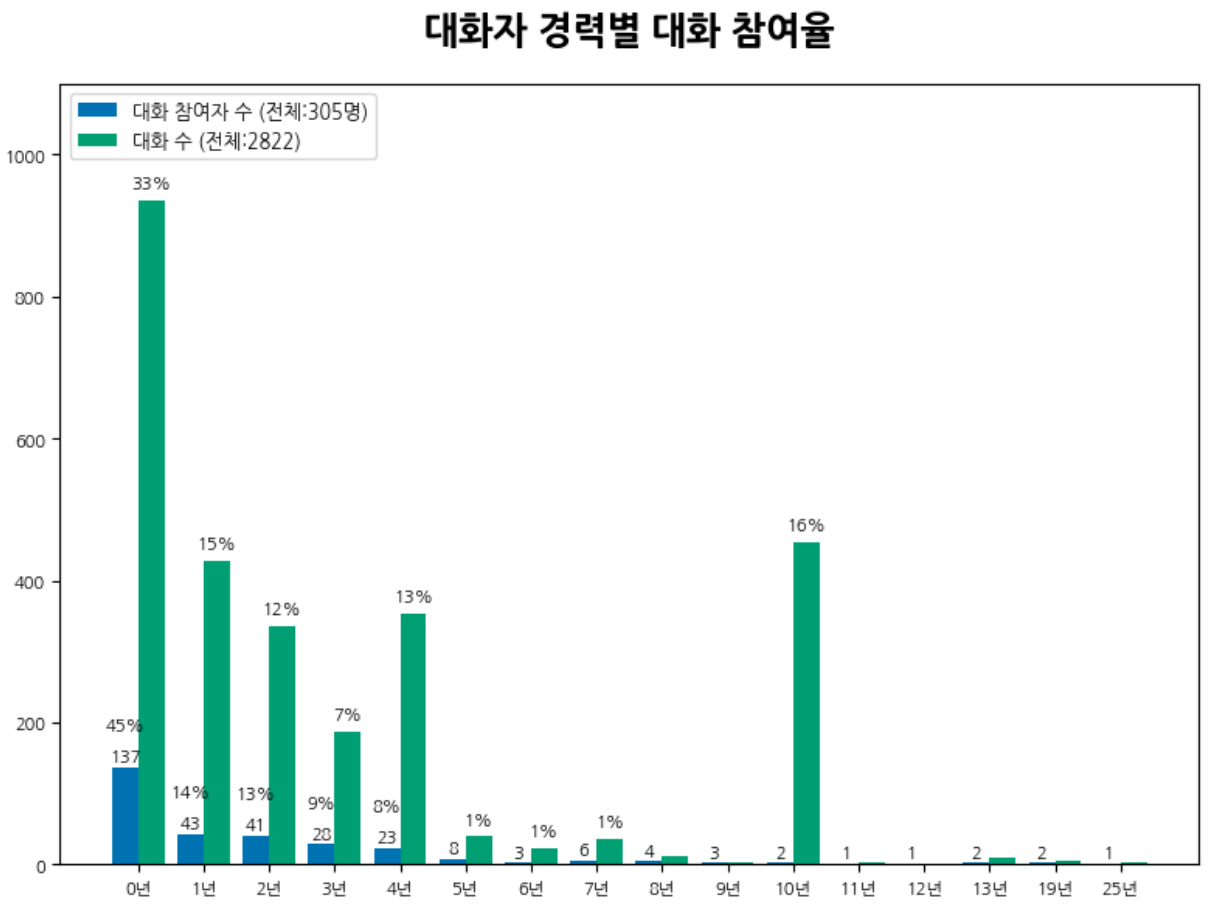
- 채팅수가 많은 Top20의 채팅 수를 막대그래프로 표현



3 채팅 참여자 파악

3-5. 경력별 참여자 파악

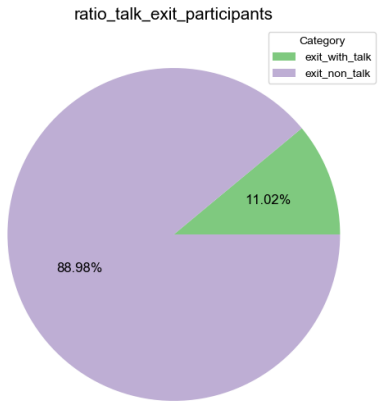
- 대학생의 경우, 0년차(신입)으로 가정
- 경력이 높아질수록 대화자의 수가 적어짐
- 전체 대화의 33%가 0년차(신입)에서 이뤄지며, 인원 대비 대화 참여율은 10년차가 가장 높게 나타남

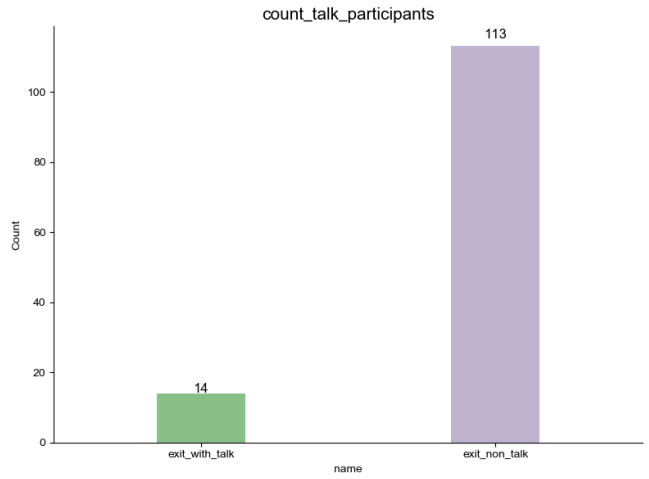


4 이탈자 채팅 여부 파악

4-1. 이탈자 채팅 횟수를 파악

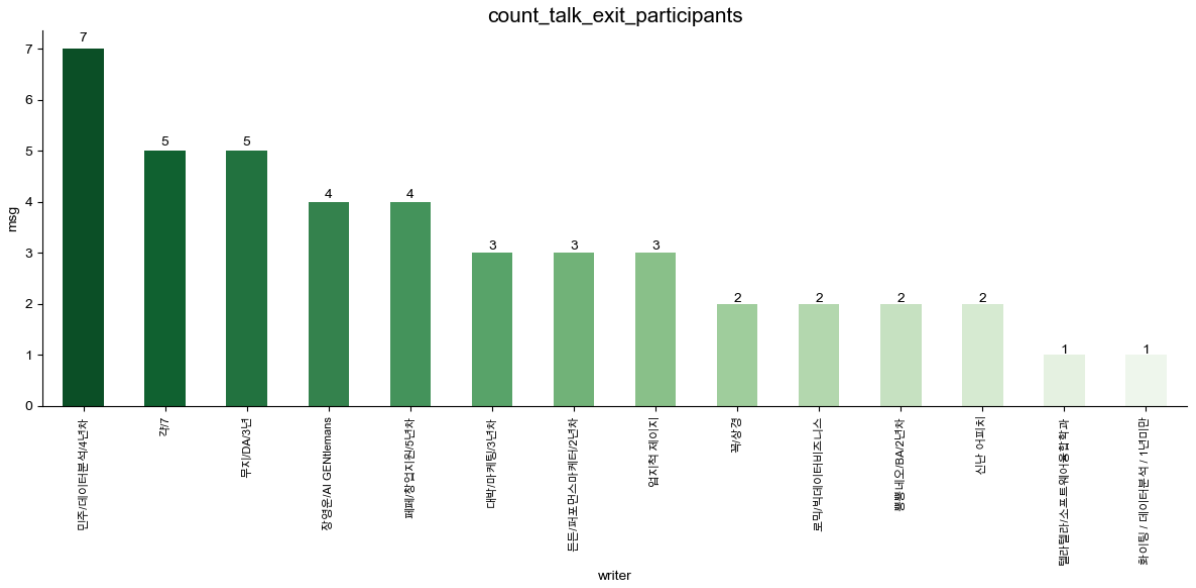
- 이탈자 중 채팅 참여자 수 및 비율을 그래프로 표현
- 이탈자 중 채팅에 참여했던 인원 파악 : 14명
- 이탈자 중 채팅 참여비율은 전체 127명 중 14명(11%)로 확인





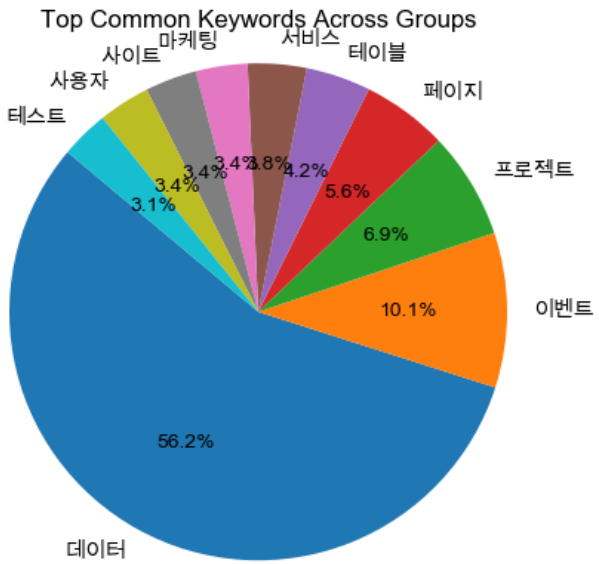
4-2. 이탈자 중 채팅 참여자의 채팅 수 확인

- 가장 많이 채팅한 이탈자의 경우 7건으로 나타남

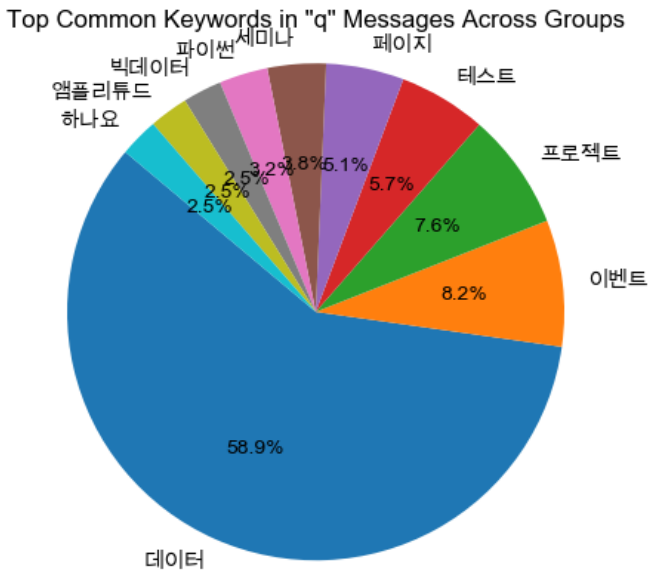


5 질문&답변 키워드 파악

채팅 내 키워드 빈도 수 상위 10개



질문자의 최초 질문 키워드 빈도 수 상위 10개



- '데이터', '이벤트', '프로젝트', '페이지' 단어의 경우 전체 채팅과 질문자의 채팅에서 두드러지게 나타남
- 질문&답변을 구분하는 키워드로는 적절하지 않다고 판단

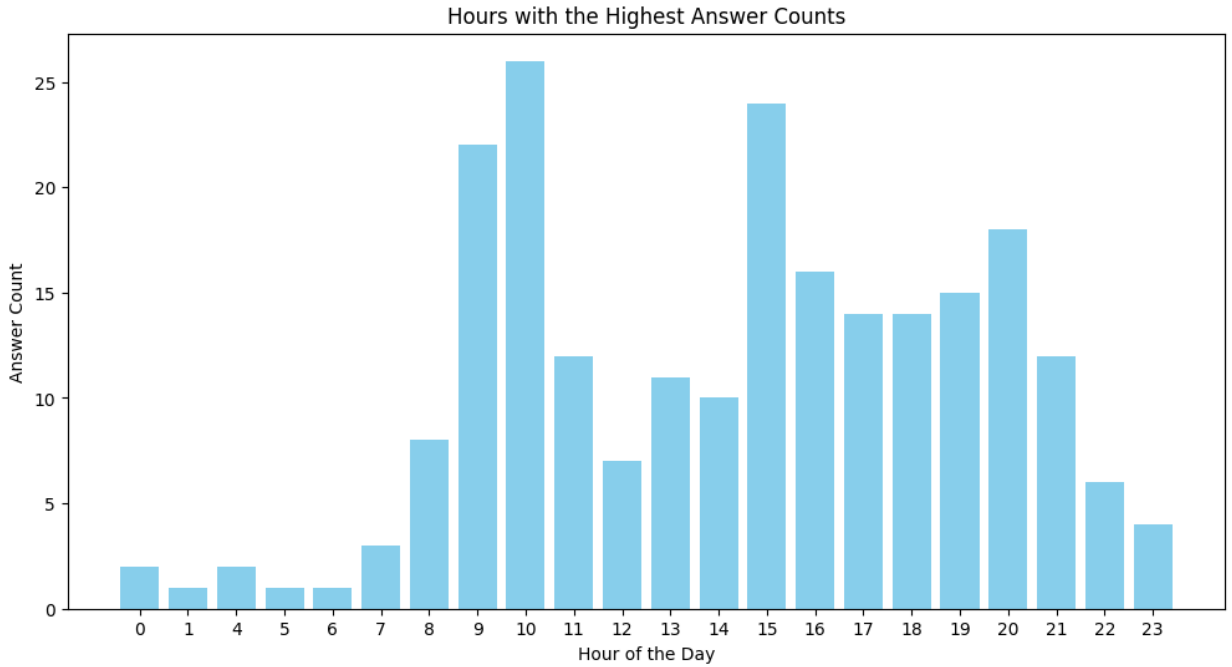
Exploratory Data Analysis

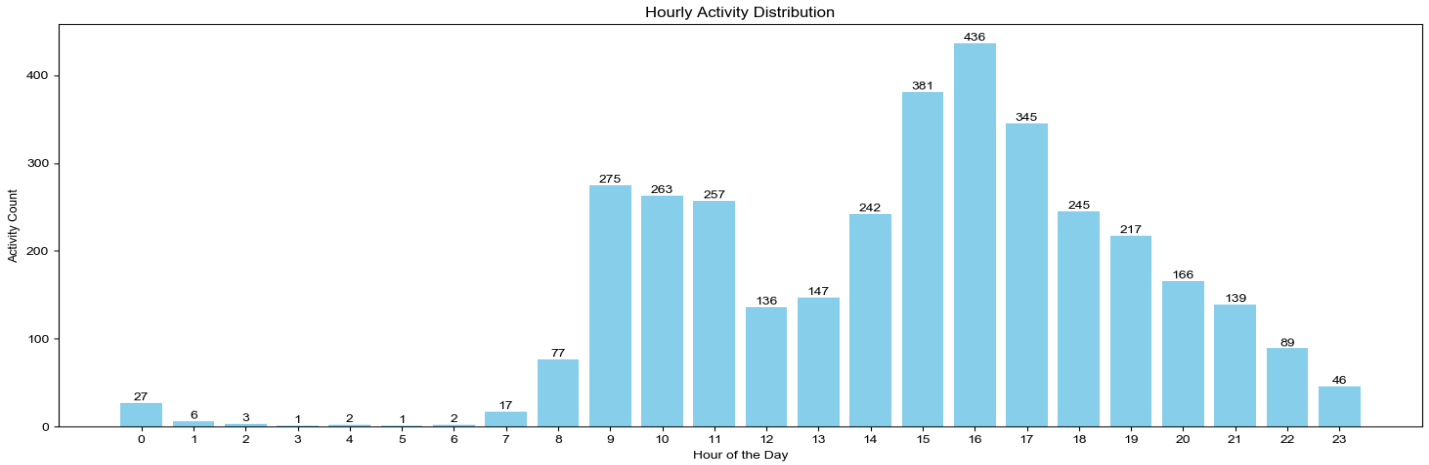
5 질문&답변 키워드 파악

5-3. 시간대별 답변 수

- 활동시간(9시~20시) 내 일반적으로 답변이 활발히 일어남
- 10시, 15시, 9시에 가장 많은 답변을 받음

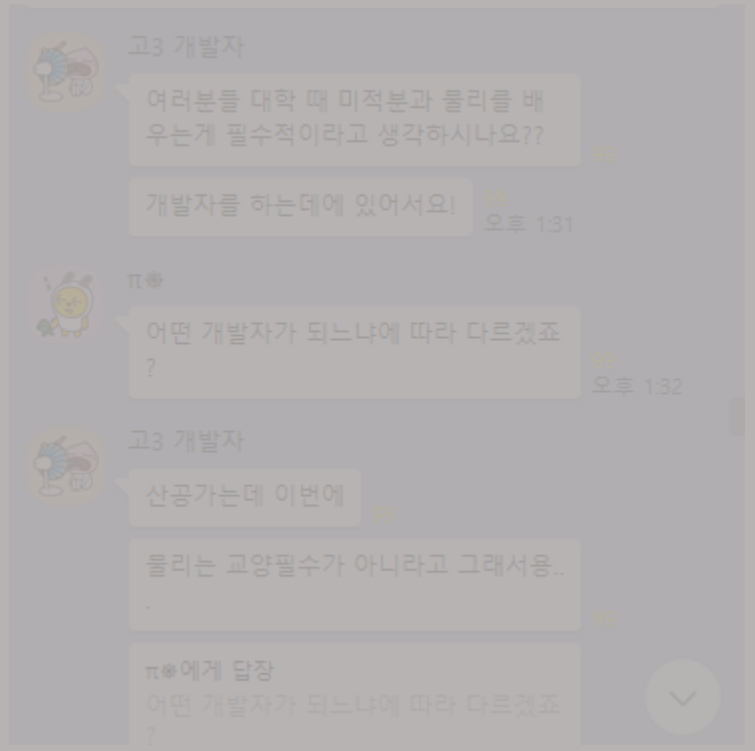
- 채팅 활동량과 답변 수를 비교하였을 때, 오전 10시, 오후 9시 이후(약 10%)를 차지





Application with Machine Learning

APPLICATION 01. 질문 구분 및 키워드 선정



무슨 주제로 이야기를 할까?

언제 이야기를 할까?

어떤 키워드로 질문을 하면 좋을까?

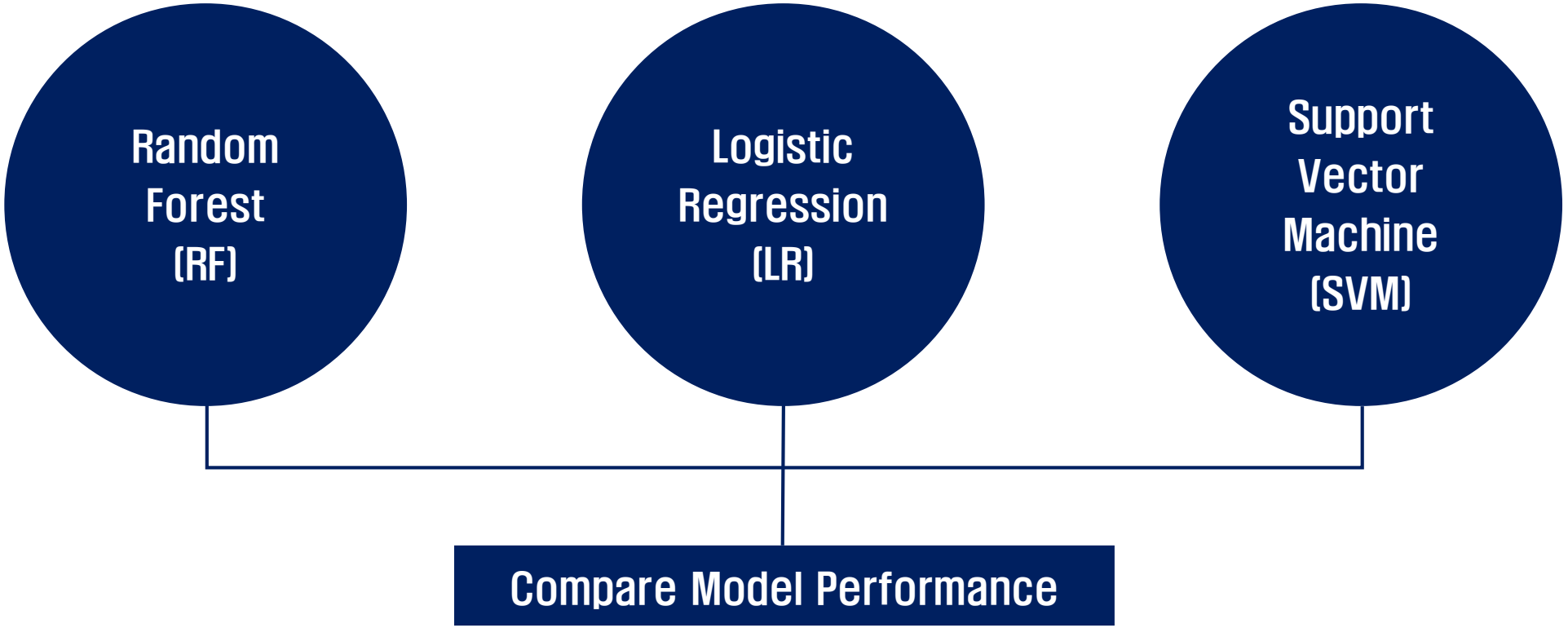
어떤 키워드를 내세워야 알맞은 답변을 받을 수 있을까?

APPLICATION 01. 질문 구분 및 키워드 선정

GOAL

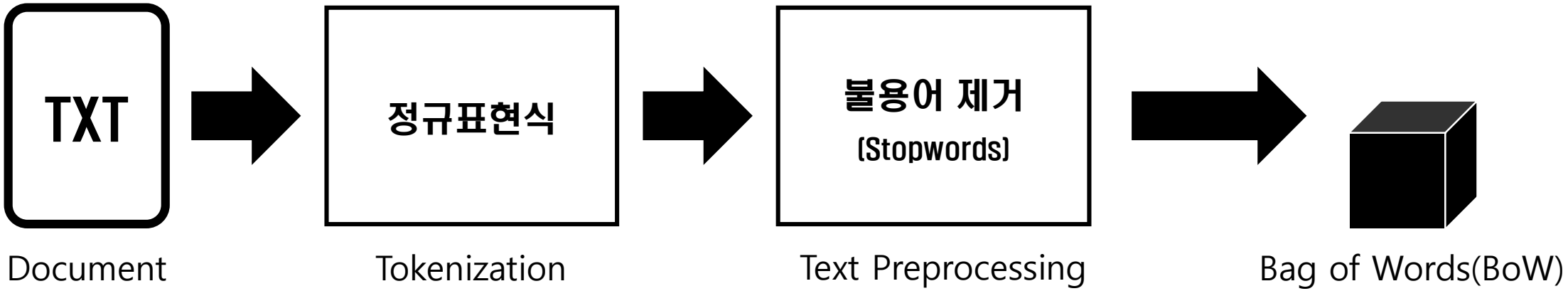
질문 구분 및 키워드 선정

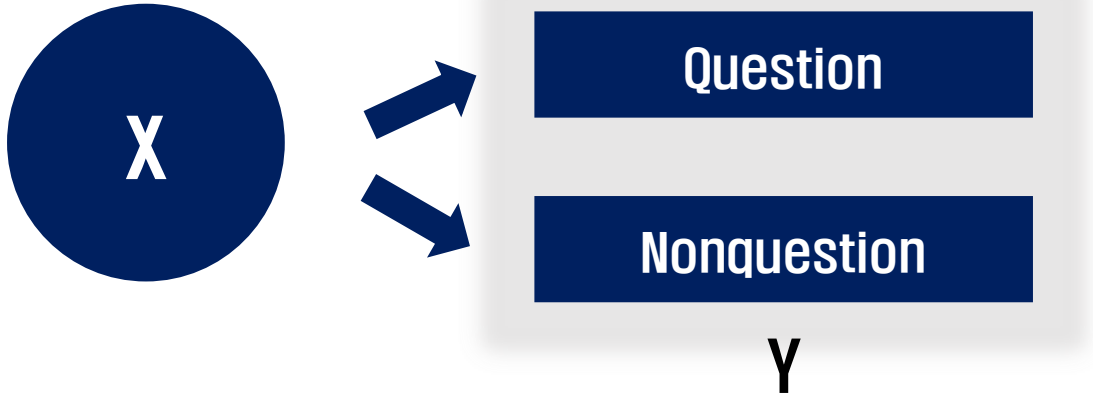
ML Model



APPLICATION 01. 질문 구분 및 키워드 선정

Preprocessing





APPLICATION 01. 질문 구분 및 키워드 선정

1 Logistic Regression

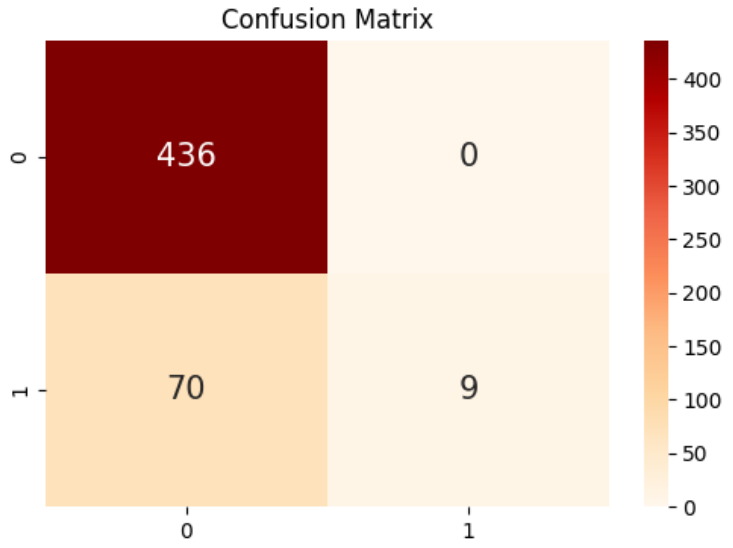
```
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# fit in training set
lr = LogisticRegression(random_state = 0)
lr.fit(x_train, y_train)

# predict in test set
y_pred = lr.predict(x_test)
```

모델 학습 결과

- Accuracy : 0.86
- Precision : 1.00
- Recall : 0.11
- F1 score : 0.20

Confusion Matrix

- 모델 평가결과를 살펴보면, 모델이 지나치게 질문이 아닌 것(0)으로만 예측하는 경향을 보임
- 따라서 질문이 아닌 것을 잘 예측하지만, 질문인 것에 대한 예측 정확도가 매우 낮게 나타남
- **샘플데이터의 클래스 불균형**으로 인한 문제로 보임



APPLICATION 01. 질문 구분 및 키워드 선정

1 Logistic Regression

1:1 sampling

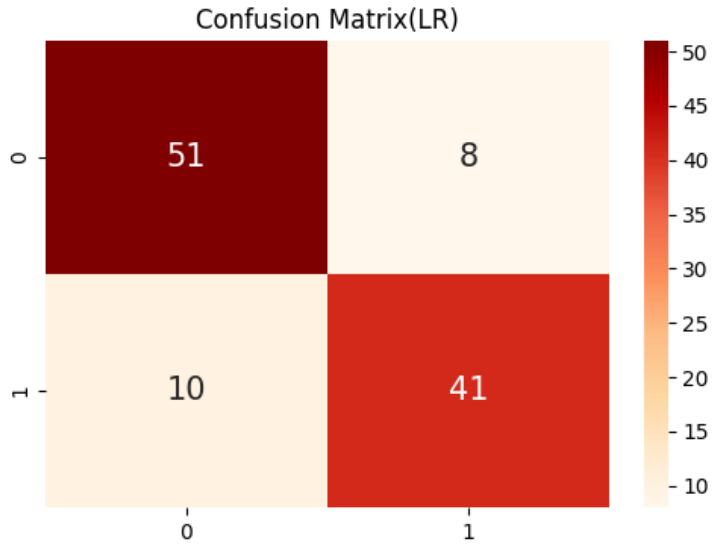
```
question_random_idx = main_df[main_df['qna_cnt']==1].sample(275, random_state=12).index.tolist()
nonquestion_random_idx = main_df[main_df['qna_cnt']==0].sample(275, random_state=12).index.tolist()
```

모델 학습 결과

- Accuracy : 0.84
- Precision : 0.84
- Recall : 0.80
- F1 score : 0.82

Confusion Matrix

- 모델이 '질문' 케이스와 '질문이 아닌' 케이스를 모두 적당히 잘 맞춘 것을 확인

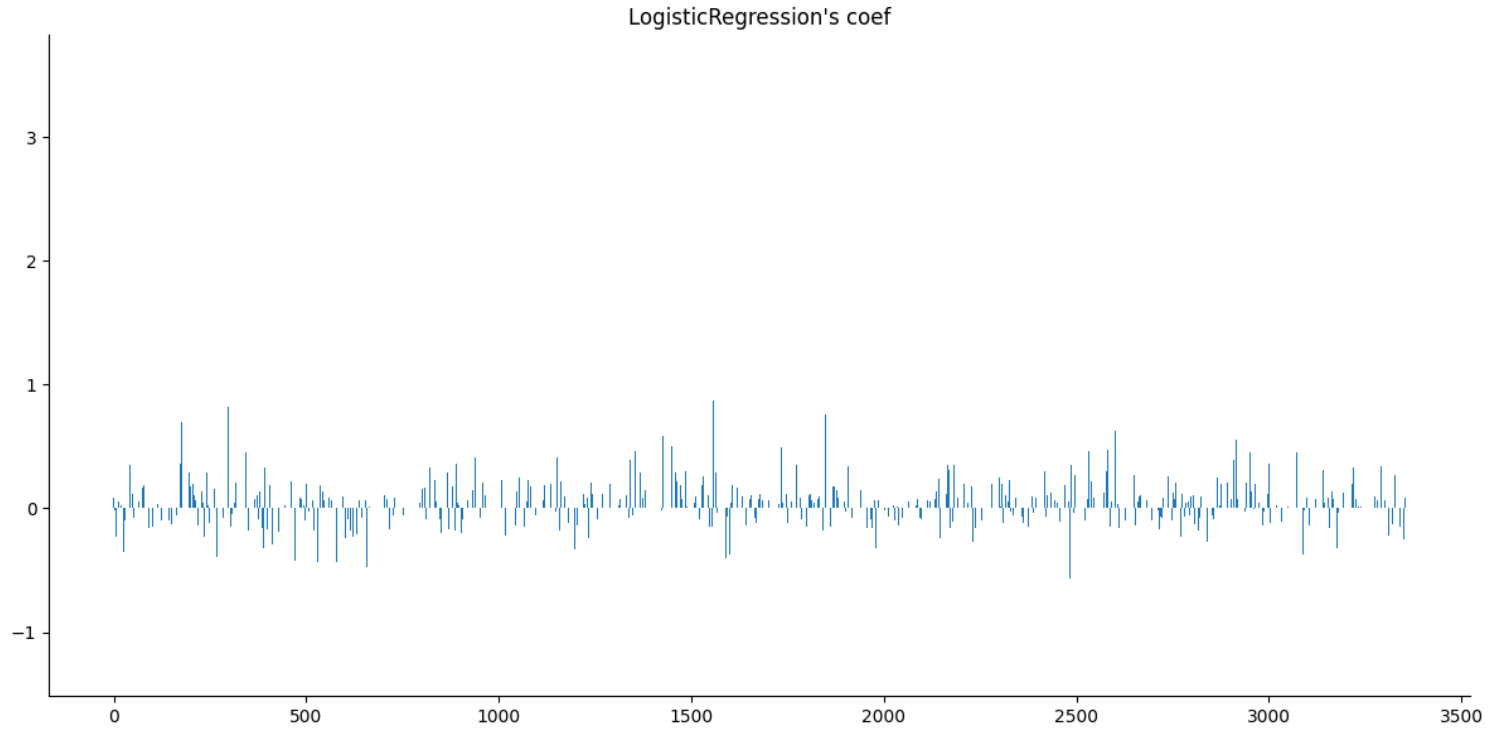


APPLICATION 01. 질문 구분 및 키워드 선정

1 Logistic Regression

키워드 분석

- 학습된 Logistic Regression 모델을 이용하여 키워드 추출
- 질문과 질문이 아닌 키워드를 추출하기 위해 먼저 Logistic Regression 모델에 각 단어의 coeficient를 확인



APPLICATION 01. 질문 구분 및 키워드 선정

1 Logistic Regression

키워드 분석

- 단어로 변환하여 '질문 키워드 리스트'와 '질문이 아닌 키워드 리스트'의 Top 20 단어를 출력

```
혹시 3.580476374668756
질문 1.6009903762790445
관련 1.4639272389366156
경험 1.134394661341066
보신 1.1023226229582899
방법 1.0307742538659916
데이터 0.9669702143891123
하나 0.892864972715186
분석 0.8834642886872529
수집 0.869688140534368
신분 0.8562816141870633
해결 0.8468256776460736
구매 0.8161244812775751
확인 0.7989685832960789
업무 0.7576632605699466
가요 0.7454699799213546
사용자 0.7201685679134158
경우 0.6955196901554933
정도 0.6931792164007218
하나요 0.6633208117754842
```

```
답변 -1.2743161354900239
아하 -0.9162173768039833
한번 -0.6989098458916249
말씀 -0.6566963253429469
부서 -0.627029693509496
감사 -0.6186245144217044
입사 -0.5858977862869876
생각 -0.5755582101941595
정말 -0.5640300245183766
발생 -0.5584680270016987
나중 -0.4974498114685263
도메인 -0.4736783597860471
보시 -0.4595440788194462
액션 -0.4587961748470687
합격 -0.45318433918171697
때문 -0.4512665717490886
인지 -0.4508407197331382
이직 -0.4501492882786481
다만 -0.4357251958164099
당장 -0.4322619548486085
```

APPLICATION 01. 질문 구분 및 키워드 선정

2 Support Vector Machine(SVM)

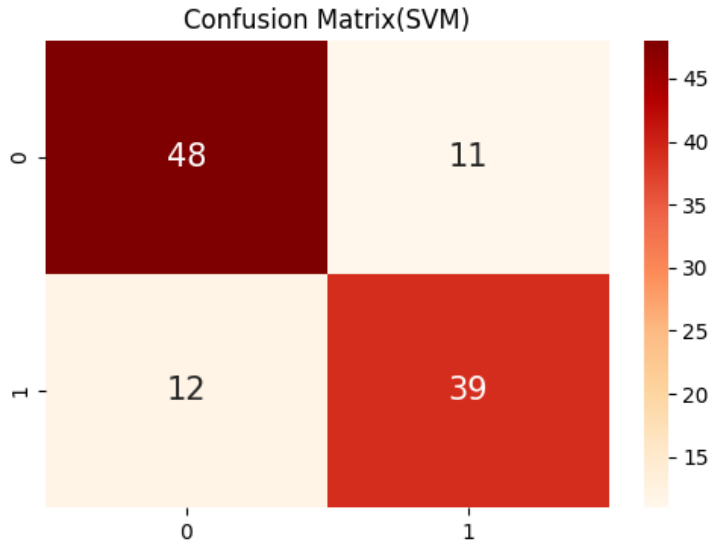
```
# 모델 2 : SVM

from sklearn.svm import SVC
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score

# fit in training set
svm = SVC(kernel='linear', C=1.0)
svm.fit(x_train, y_train)

# predict in test set
y_pred = svm.predict(x_test)
```

모델 학습 결과

- Accuracy : 0.79
- Precision : 0.78
- Recall : 0.76
- F1 score : 0.77

Confusion Matrix

- 모델이 '질문' 케이스와 '질문이 아닌' 케이스를 모두 적당히 잘 맞춘 것을 확인



APPLICATION 01. 질문 구분 및 키워드 선정

3 Random Forest (RF)

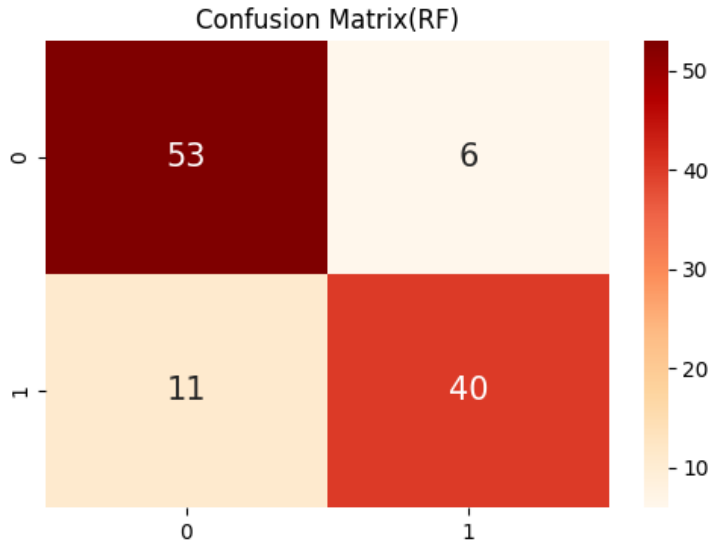
```
# 모델 3 : RF
from sklearn.ensemble import RandomForestClassifier

# Random Forest 모델 생성 및 훈련
rf_model = RandomForestClassifier(n_estimators=100, random_state=0)  # 100개의 트리를 사용
rf_model.fit(x_train, y_train)

# 테스트 데이터에 대한 예측
y_pred = rf_model.predict(x_test)
```

모델 학습 결과

- Accuracy : 0.85
- Precision : 0.87
- Recall : 0.78
- F1 score : 0.82

Confusion Matrix

- 모델이 '질문' 케이스와 '질문이 아닌' 케이스를 모두 적당히 잘 맞춘 것을 확인



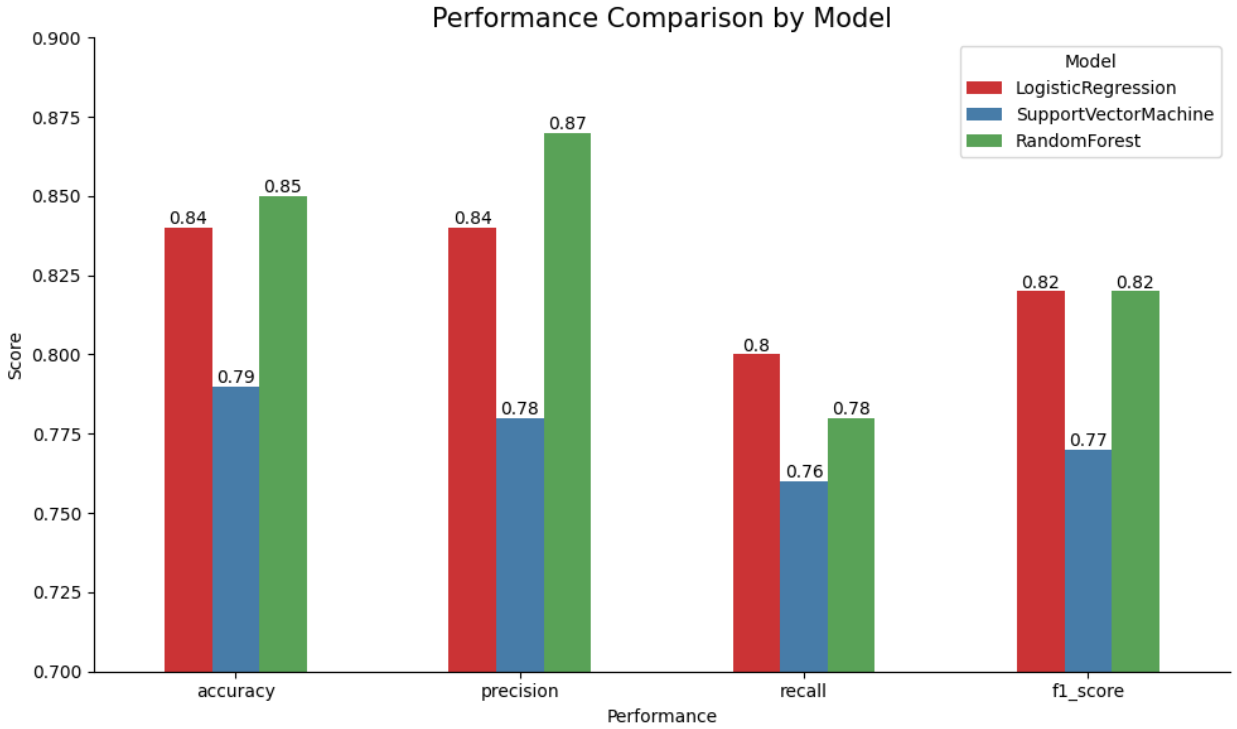
APPLICATION 01. 질문 구분 및 키워드 선정

Performance Analysis

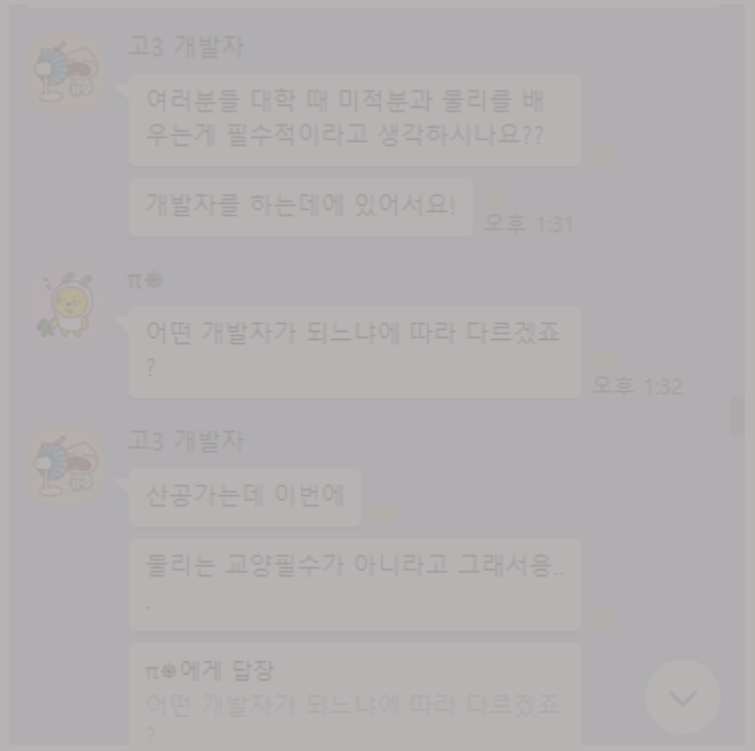
모델 비교

- 3개의 선정 모델 중 RF(Random Forest)모델이 가장 높은 성능으로 나타남

	Performance	LogisticRegression	SupportVectorMachine	RandomForest
0	accuracy	0.84	0.79	0.85
1	precision	0.84	0.78	0.87
2	recall	0.80	0.76	0.78
3	f1_score	0.82	0.77	0.82



APPLICATION 02. 시간대 별 응답률 예측



무슨 주제로 이야기를 할까?

언제 이야기를 할까?

어떤 키워드로 질문을 하면 좋을까?

어떤 키워드를 내세워야 알맞은 답변을 받을 수 있을까?

APPLICATION 02. 시간대 별 응답률 예측

GOAL

언제, 어떤 키워드로 질문하면 응답받을 수 있을까?

Preprocessing

Document → Tokenization → Text Preprocessing → Bag of Words(BoW)

ML Model

APPLICATION 02. 시간대 별 응답률 예측

1 Naive Bayes Classifier

```python
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.naive_bayes import MultinomialNB

def predict_response_probability(model, vectorizer, hour, keyword):
    # 입력 데이터를 벡터로 변환
    input_data = pd.DataFrame({'hour': [hour], 'msg': [keyword]})
    input_vectorized = vectorizer.transform(input_data['msg'])


    # 모델로 답변 확률 예측
    response_probability = model.predict_proba(input_vectorized)[:, 1]

    return response_probability[0]
```

모델 학습 결과

- Accuracy : 0.74

Application

- 시간, 키워드 설정하여 실제 답변받을 확률을 적용
- 답변 받을 확률 : 36%

```python
# 예측 함수를 사용하여 답변 확률 예측
hour = 20  # 예측하려는 시간
keyword = '분석에 대한 질문입니다.'  # 예측하려는 키워드
response_prob = predict_response_probability(model, vectorizer, hour, keyword)
print(response_prob)
```

Problem

- 키워드는 동일, 시간을 변경하였을 때 답변 확률 변화가 X