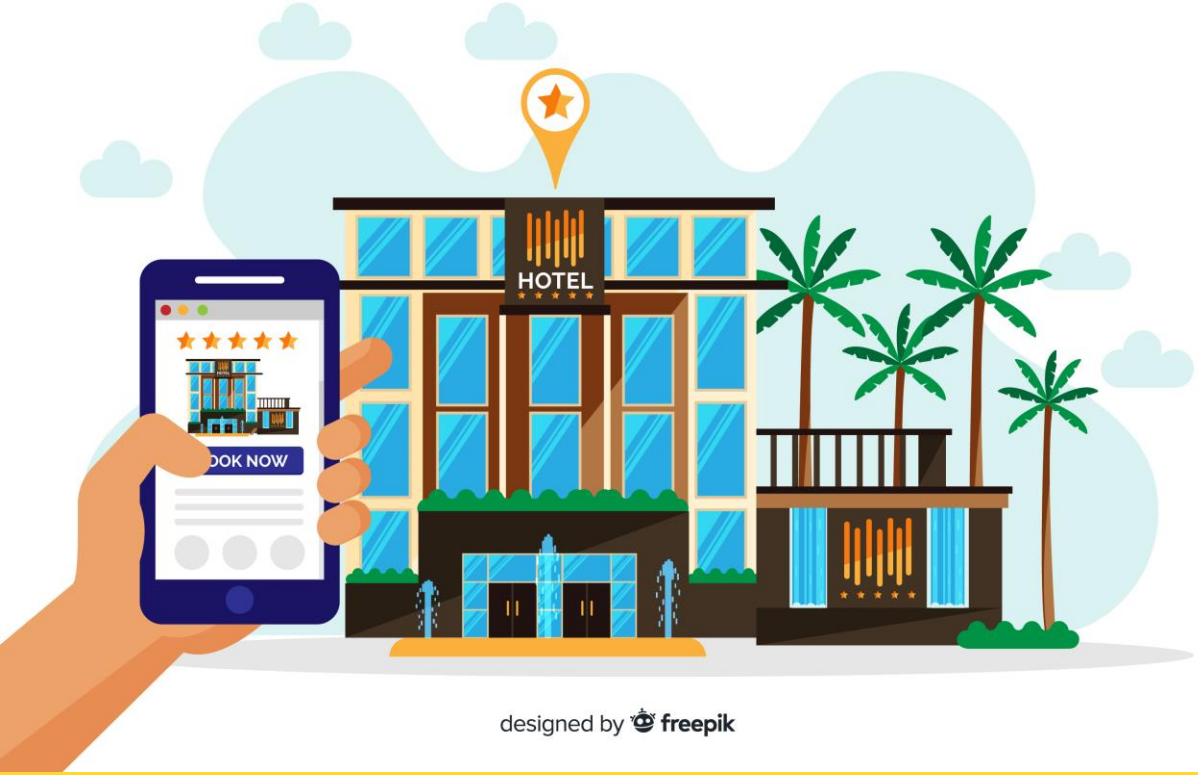
호텔 리뷰 분석



머신러닝 2팀

고지현, 정기중, 황지우

INDEX

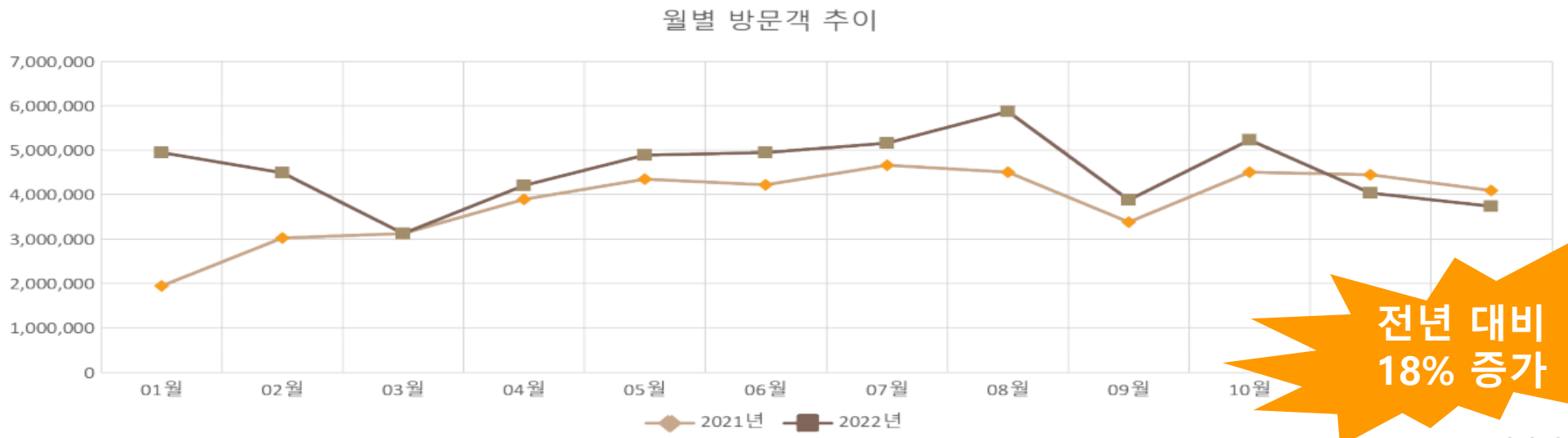
Chapter.1

주제 선정

Chapter.1

주제 선정

내국인 전체 방문객



전년 대비
18% 증가

(단위:명)

인구(명)	01월	02월	03월	04월	05월	06월	07월	08월	09월	10월	11월	12월	총합계
2021년	1,957,992	3,044,100	3,140,626	3,903,751	4,359,043	4,229,449	4,679,390	4,511,143	3,386,594	4,515,758	4,463,014	4,106,147	46,297,007
2022년	4,953,781	4,503,977	3,137,738	4,224,766	4,897,263	4,953,764	5,173,971	5,880,167	3,885,247	5,243,988	4,047,771	3,748,582	54,651,015
인구차이	2,995,789	1,459,877	-2,888	321,015	538,220	724,315	494,581	1,369,024	498,653	728,230	-415,243	-357,565	8,354,008
증감률	153%	48%	-0.1%	8.2%	12.3%	17.1%	10.6%	30.3%	14.7%	16.1%	-9.3%	-8.7%	18%

Chapter.1

주제 선정

목적

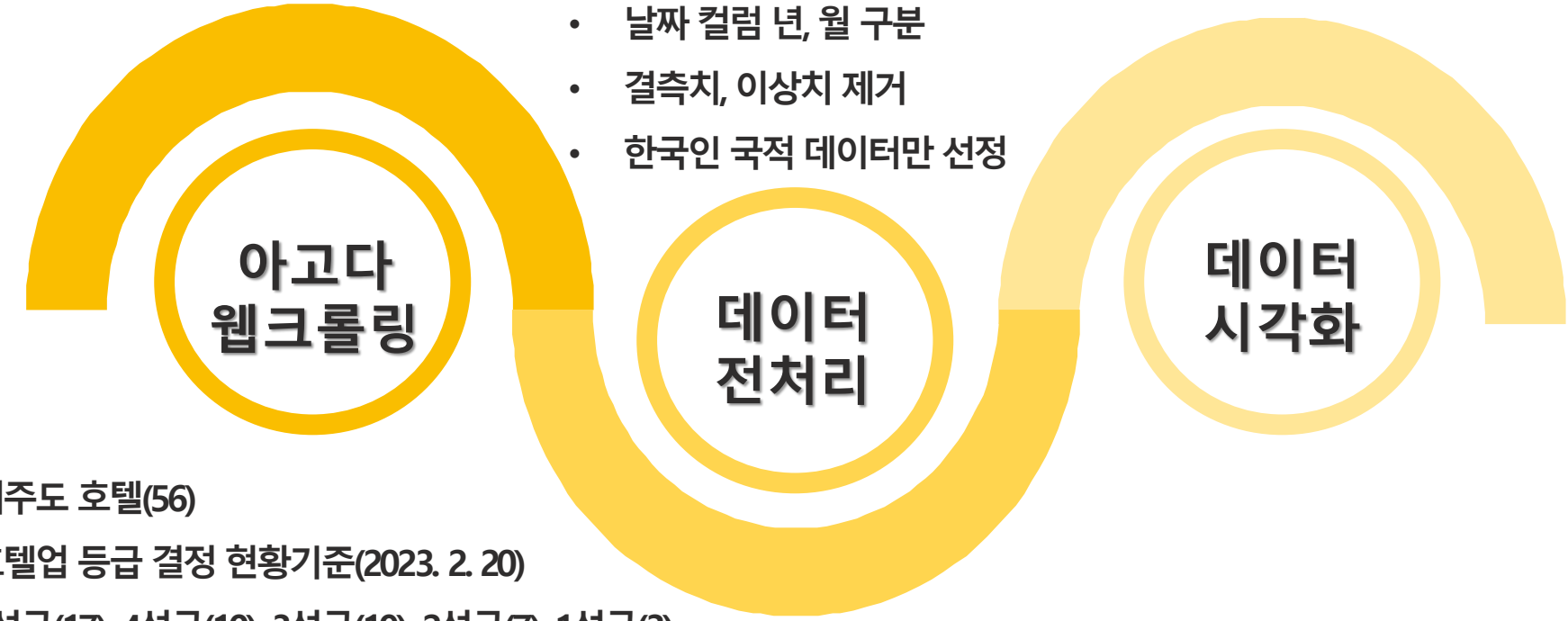
- 리뷰 키워드 이용하여 호텔 리뷰 긍정/ 부정 감성분석

목표

- 다른호텔 리뷰 평점 없이 리뷰 내용으로 긍정/ 부정 구분

Chapter.2

데이터

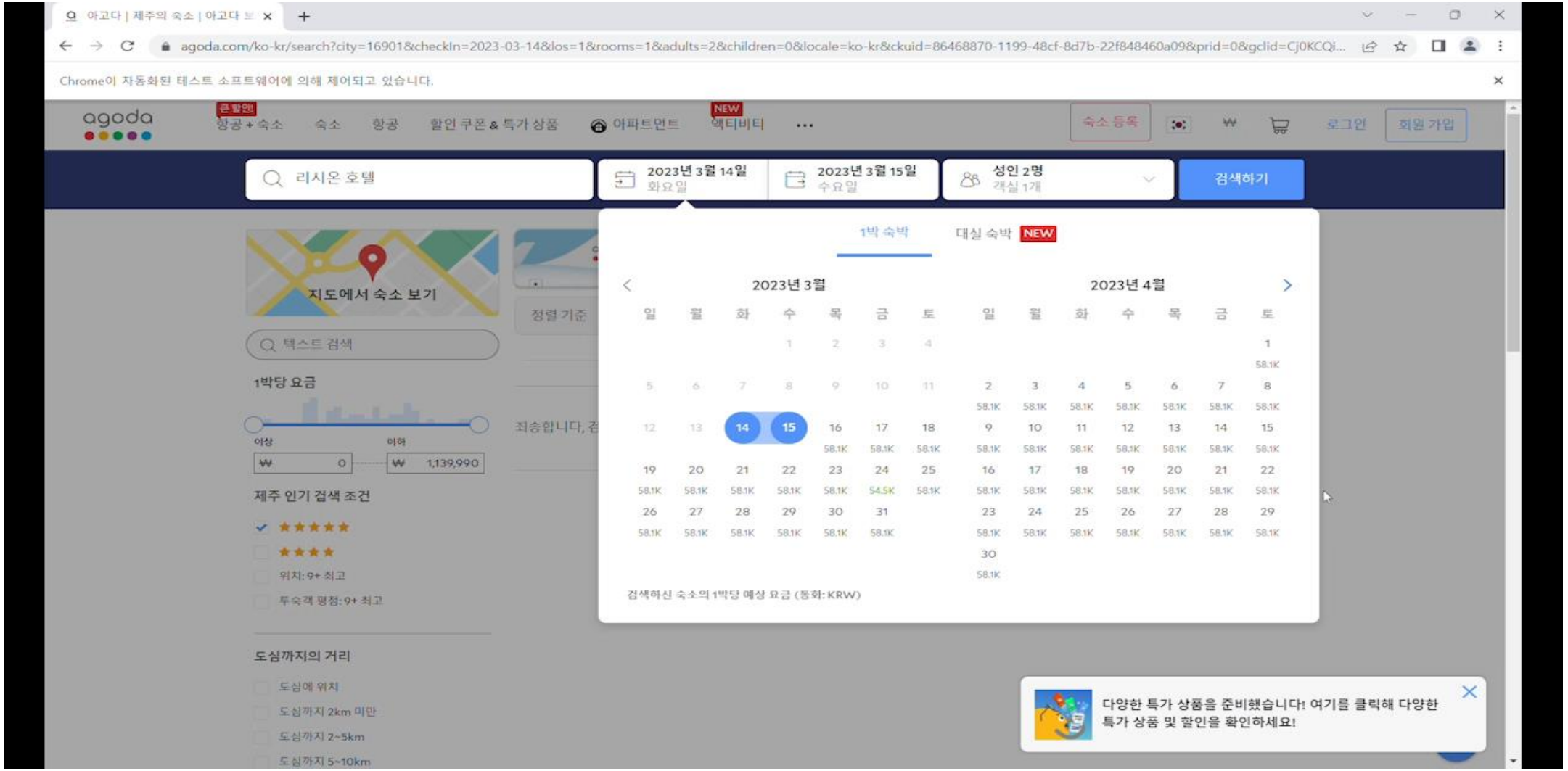


- 제주도 호텔(56)
- 호텔업 등급 결정 현황기준(2023. 2. 20)
- 5성급(17), 4성급(10), 3성급(19), 2성급(7), 1성급(3)

- 날짜 컬럼 년, 월 구분
- 결측치, 이상치 제거
- 한국인 국적 데이터만 선정

Chapter.2

데이터 수집

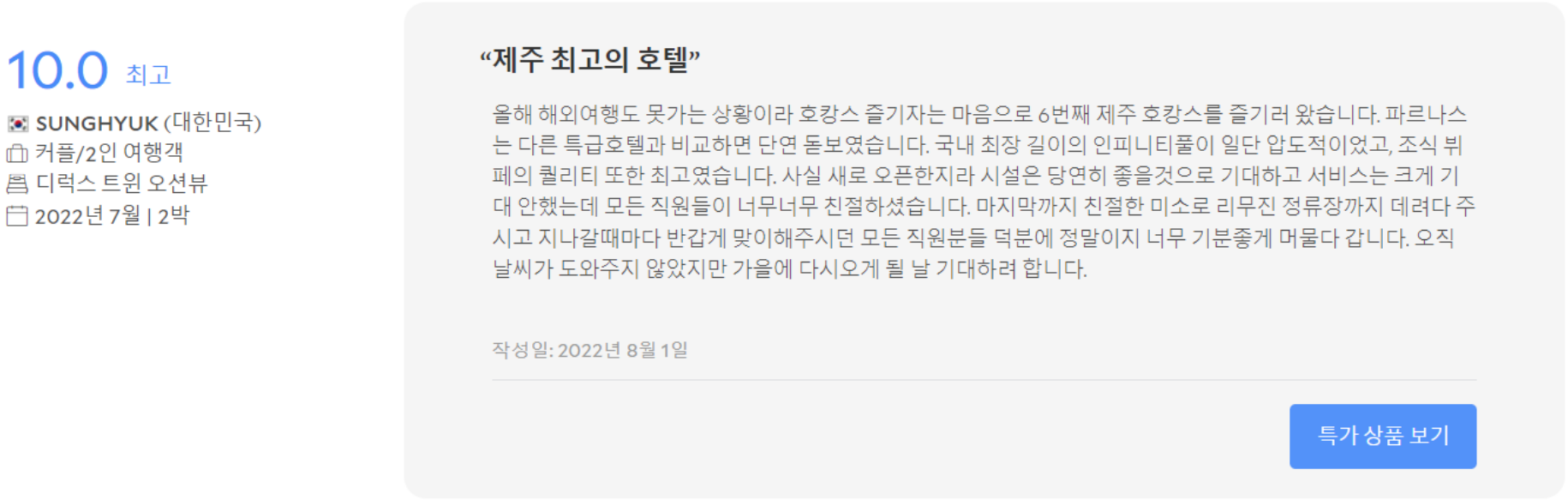


Chapter.2

데이터 - raw data

컬럼명

- 호텔이름, 평점, 국적, 맴버, 룸 타입, 날짜, 리뷰 제목, 리뷰 내용

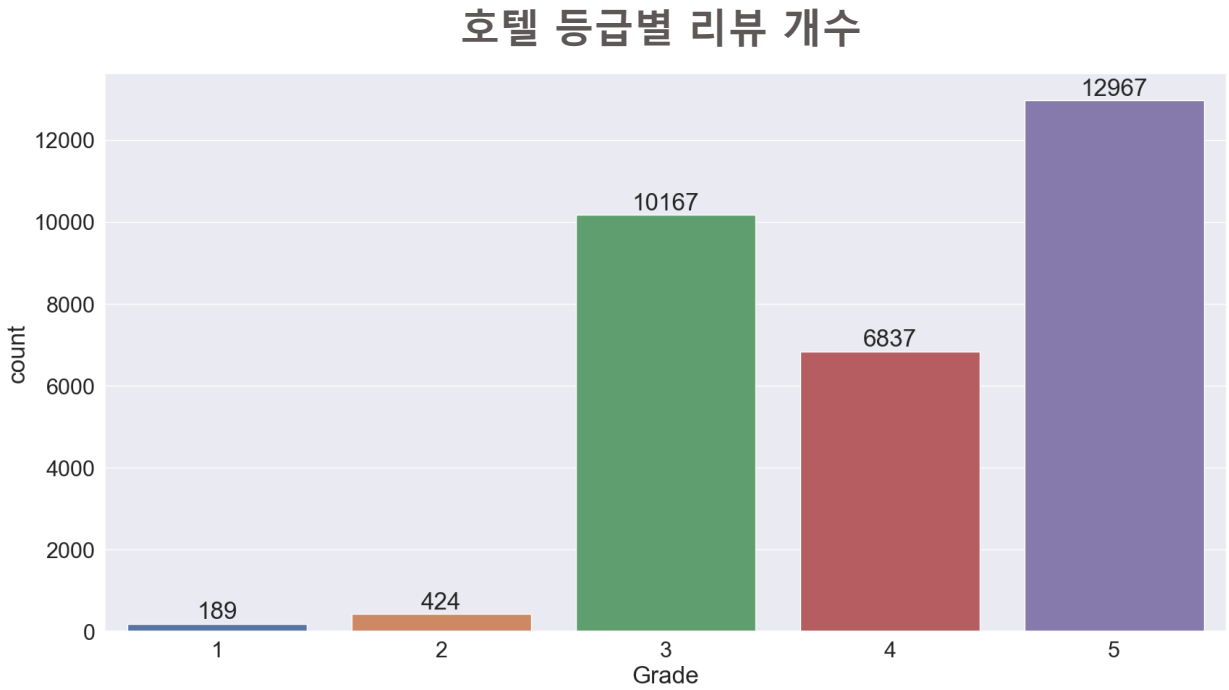


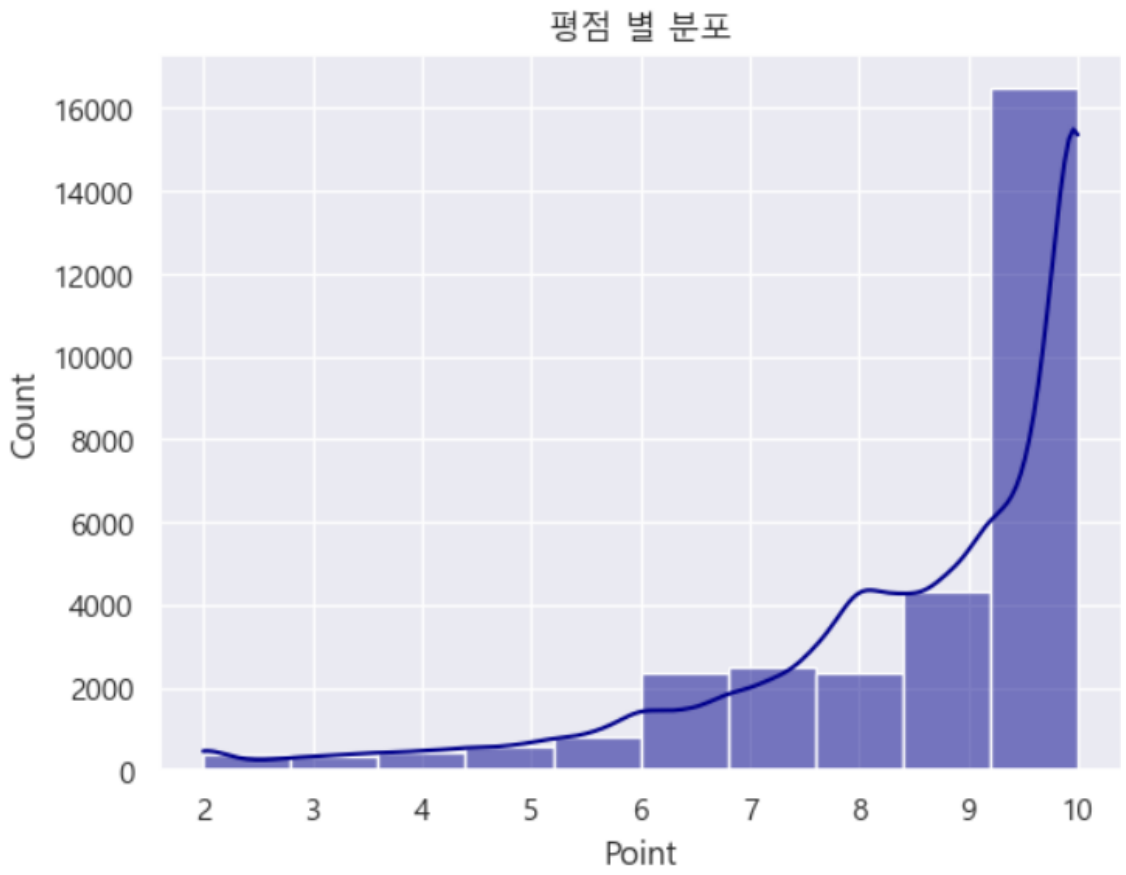
Chapter.2

데이터 - preprocessing data (30584개)

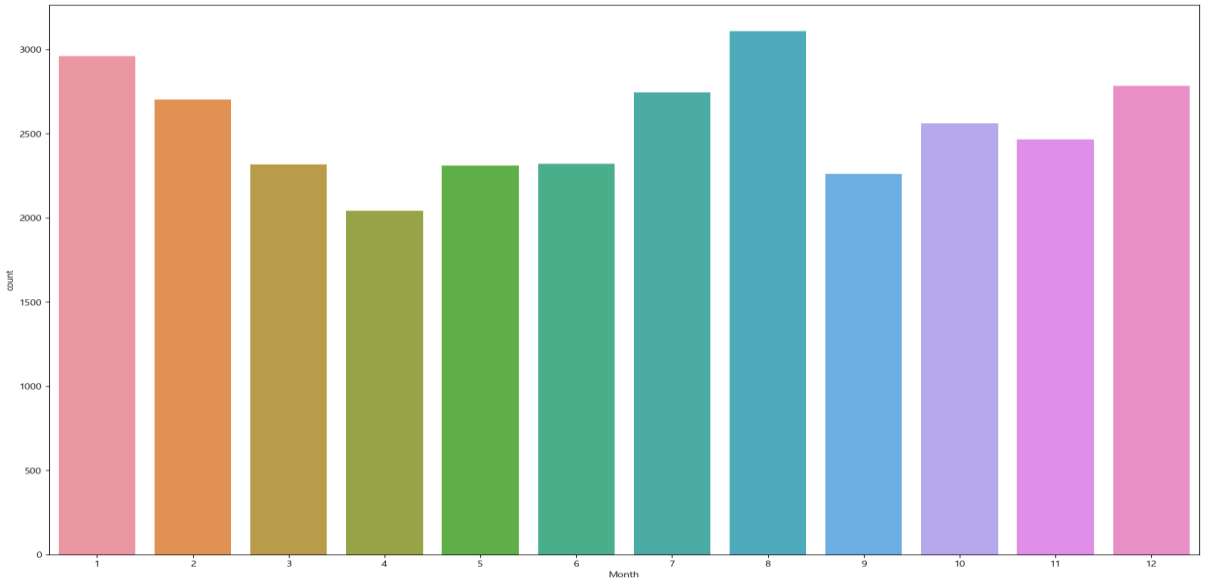
Chapter.2

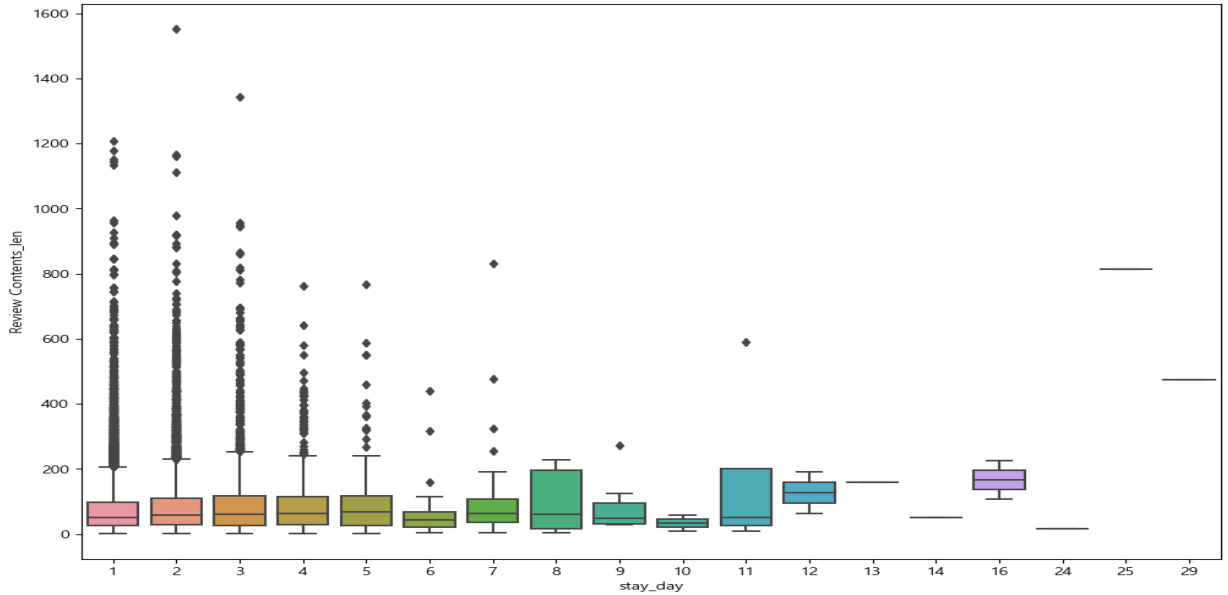
데이터 시각화

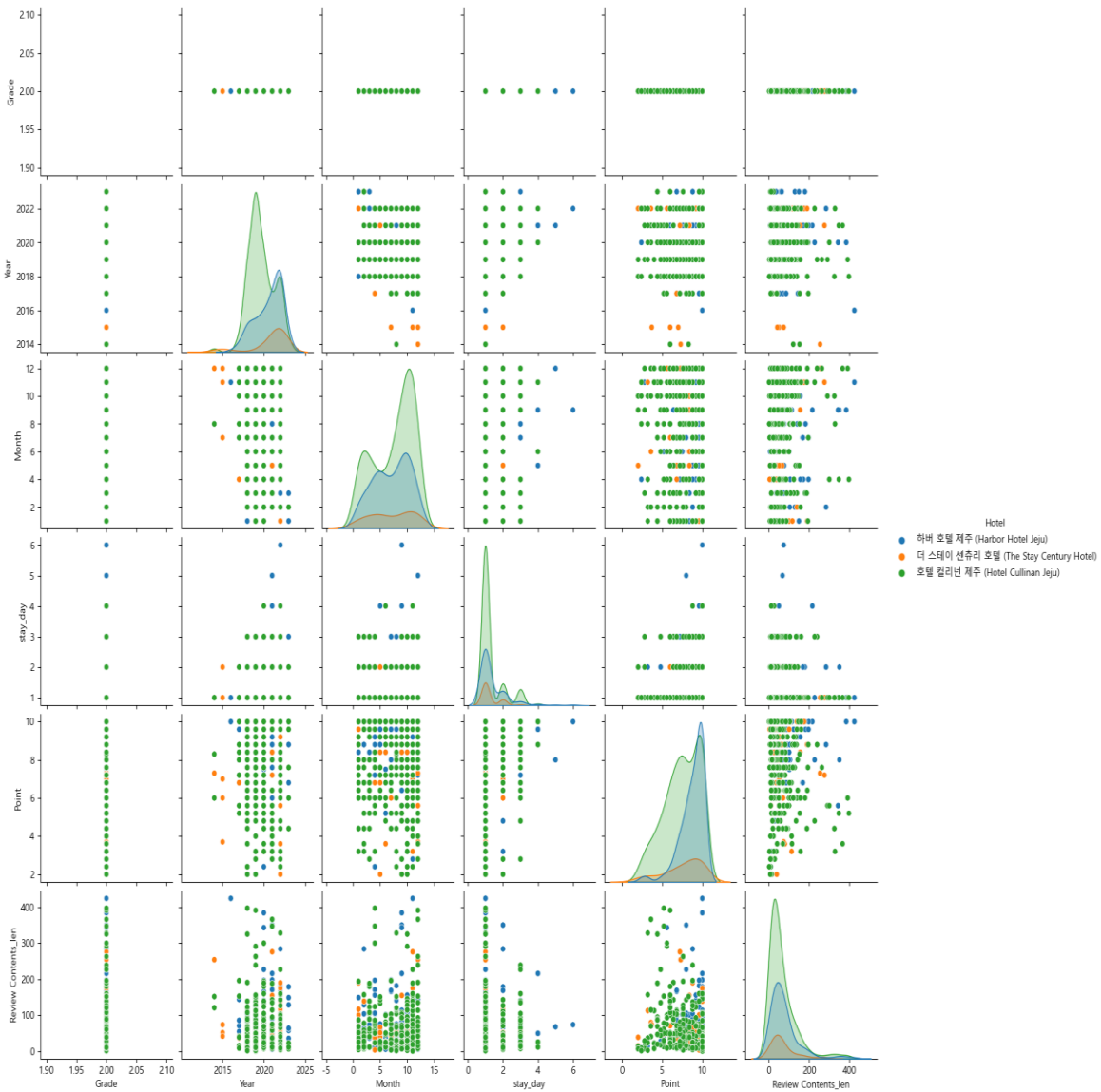




월별 호텔 리뷰 수


count
Month


Review Contents_len
stay_day


Grade
Year
Month
stay_day
Point
Review Contents_len
Hotel
하버 호텔 제주 (Harbor Hotel Jeju)
더 스테이 센츄리 호텔 (The Stay Century Hotel)
호텔 컬리넌 제주 (Hotel Cullinan Jeju)

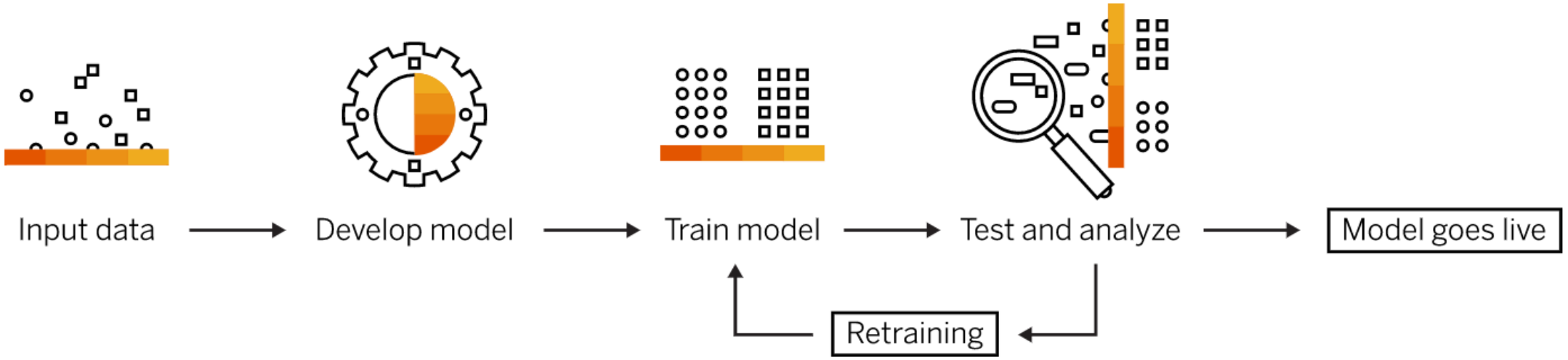
Chapter.2

EDA 결과

- 호텔등급이 높을수록 평균 평점이 높음
- 평점이나 숙박 기간, 호텔등급, 이용 인원과 후기 글자 수는 크게 연관이 없음
- 제주도는 1월과 8월에 이용객이 많음
  -> 방학 기간에 커플과 가족 여행 인원이 증가하는 것으로 추정
- 커플이나 가족은 주로 5성급 이용
- 나 홀로 여행객 및 출장 여행객은 3성급 이용
- 월별로 호텔 선호순위가 다름

Chapter.3

머신러닝



머신러닝 프로세스 작동 방식

Chapter.3

데이터 분석 - 분류

- 6~9점 사이 리뷰

주차 때 실갱이 아닌 실갱이로 기분 상함

전체적으로 만족합니다

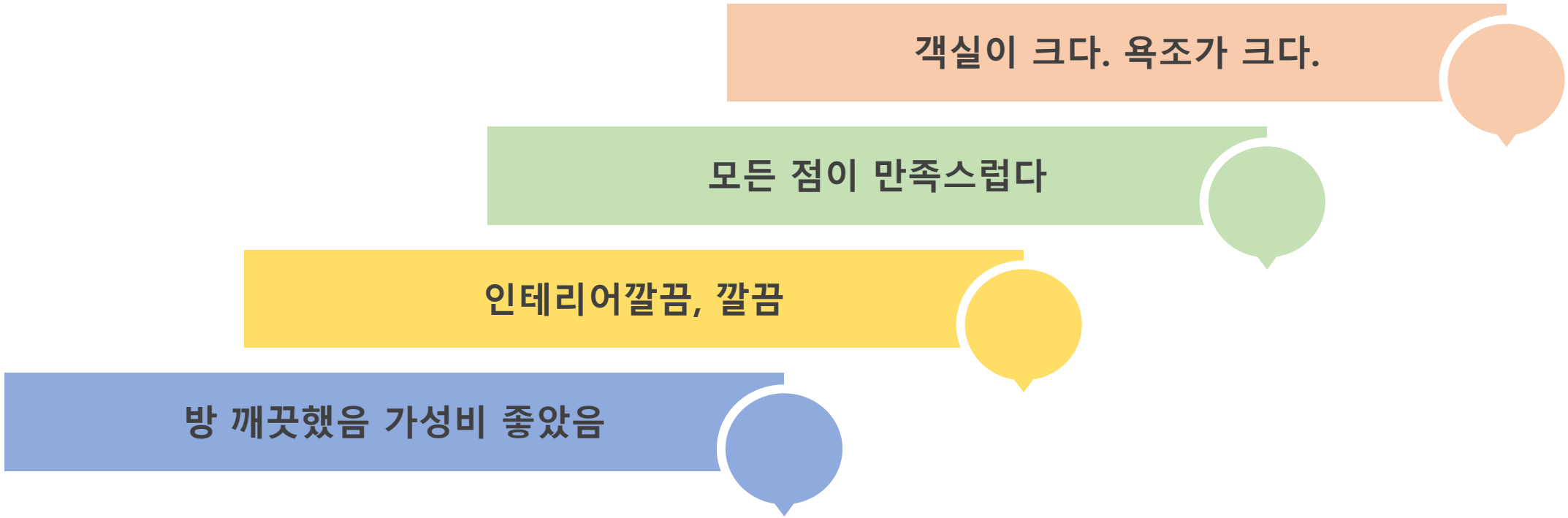
수화기에 먼지가 가득 쌓여 있음

산책 말고는 딱히 할게 없었어요

Chapter.3

데이터 분석 - 분류

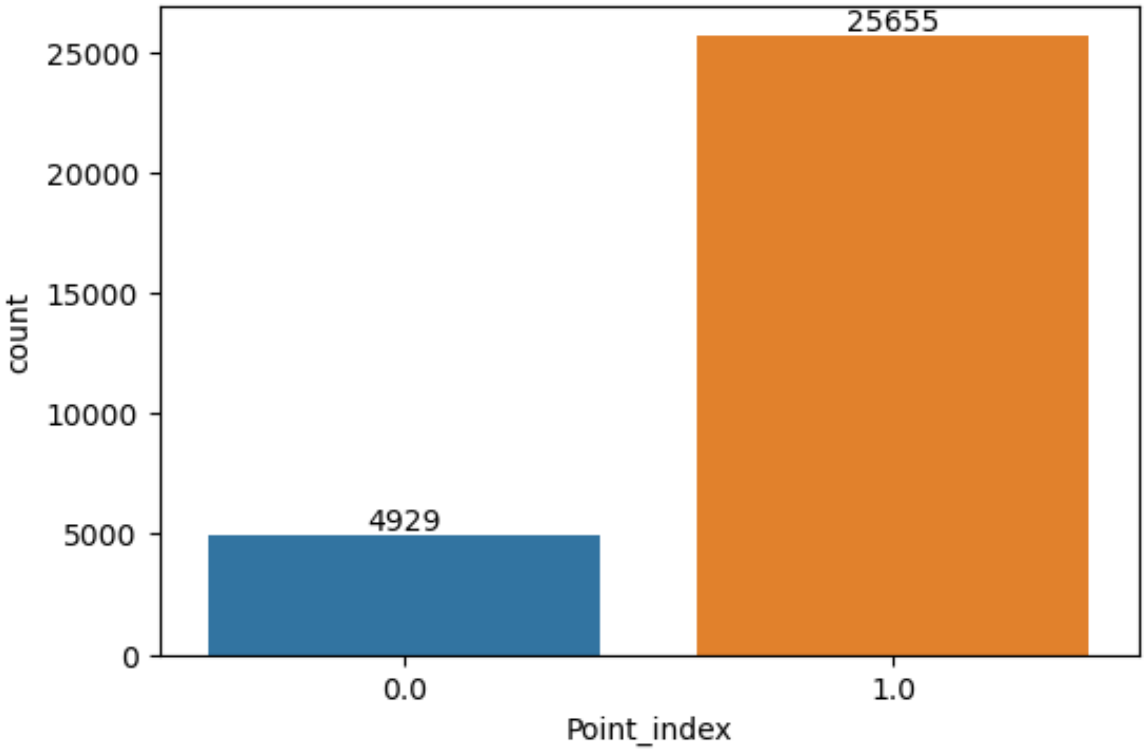
- 9점 이상



Chapter.3

데이터 분석 - 분류

2. 7점 기준 긍정/부정(형태소/ 명사)

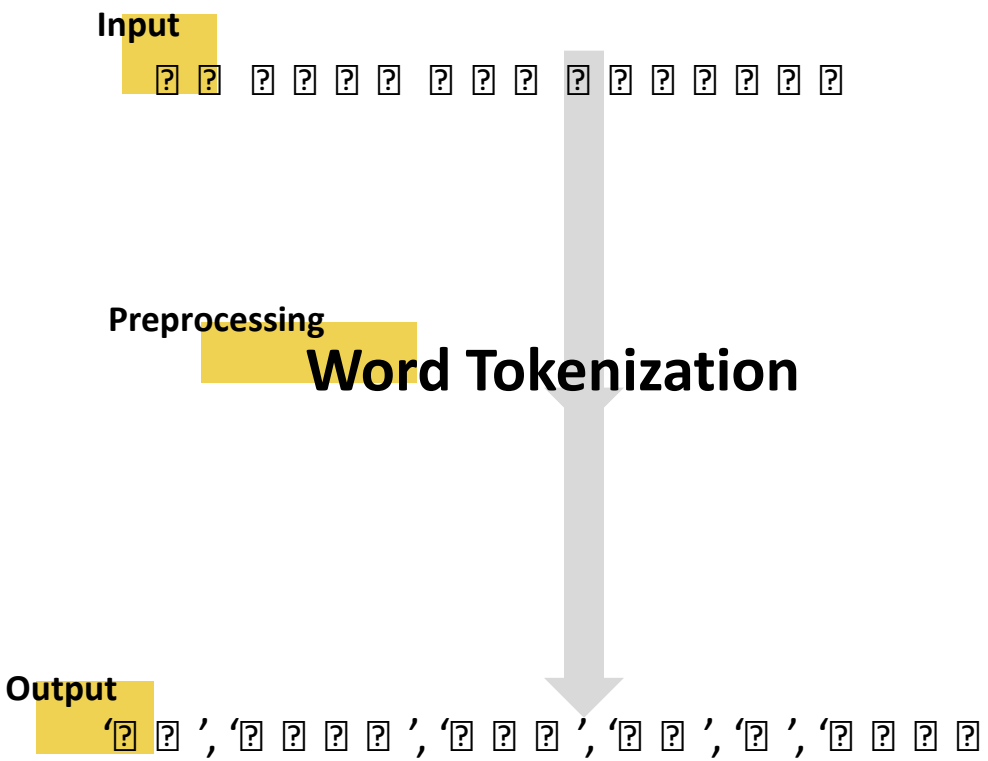


형태소 기준

명사 기준

Chapter.3

자연어 처리

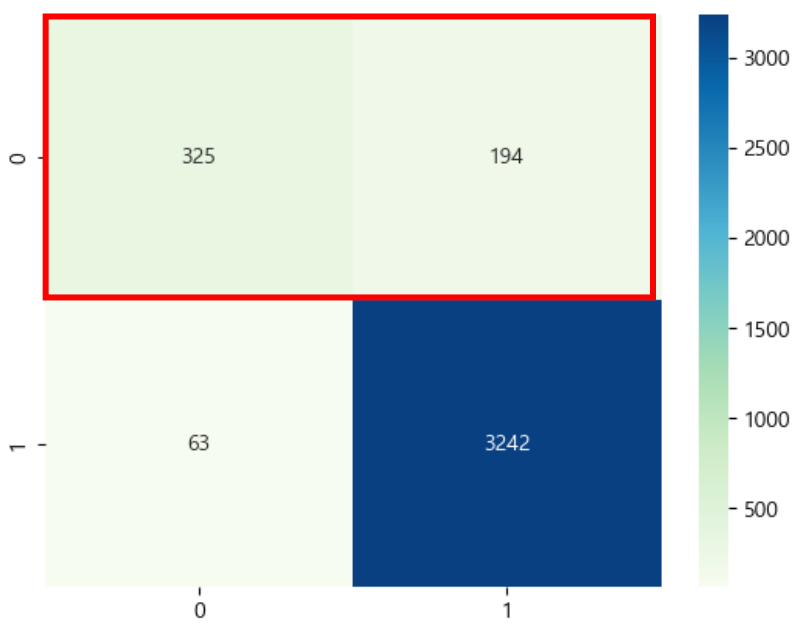


Chapter.3

모델 선정

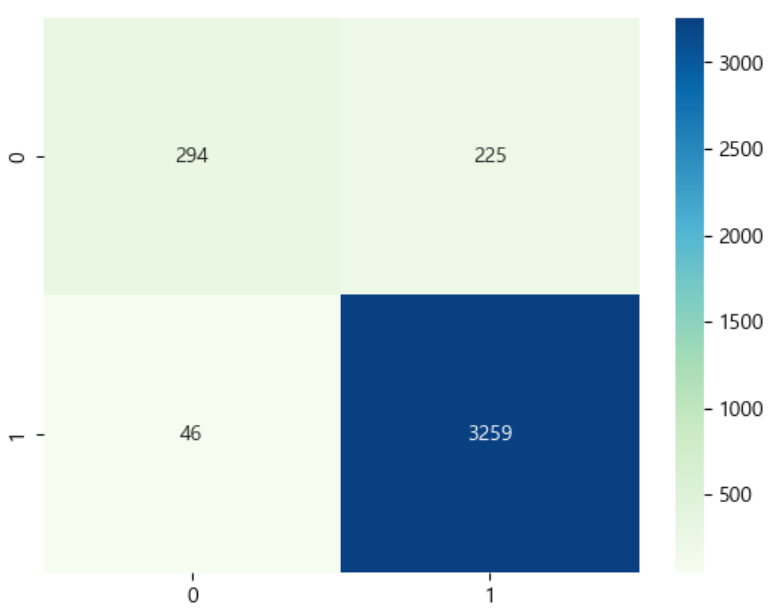
- 나이브베이즈 성능이 가장 좋으므로 **나이브베이즈 모델** 선정

나이브베이즈



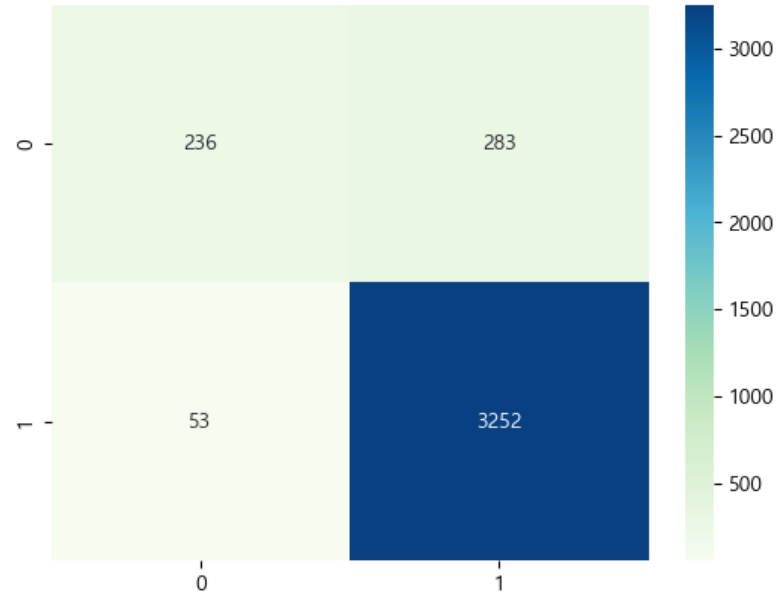
Accuracy : 0.9328, Precision : 0.9435
Recall : 0.9840, F1 : 0.9509, Auc:0.7193

로지스틱 회귀



Accuracy : 0.9291, Precision : 0.9354
Recall : 0.9861, F1 : 0.9601, Auc:0.7763

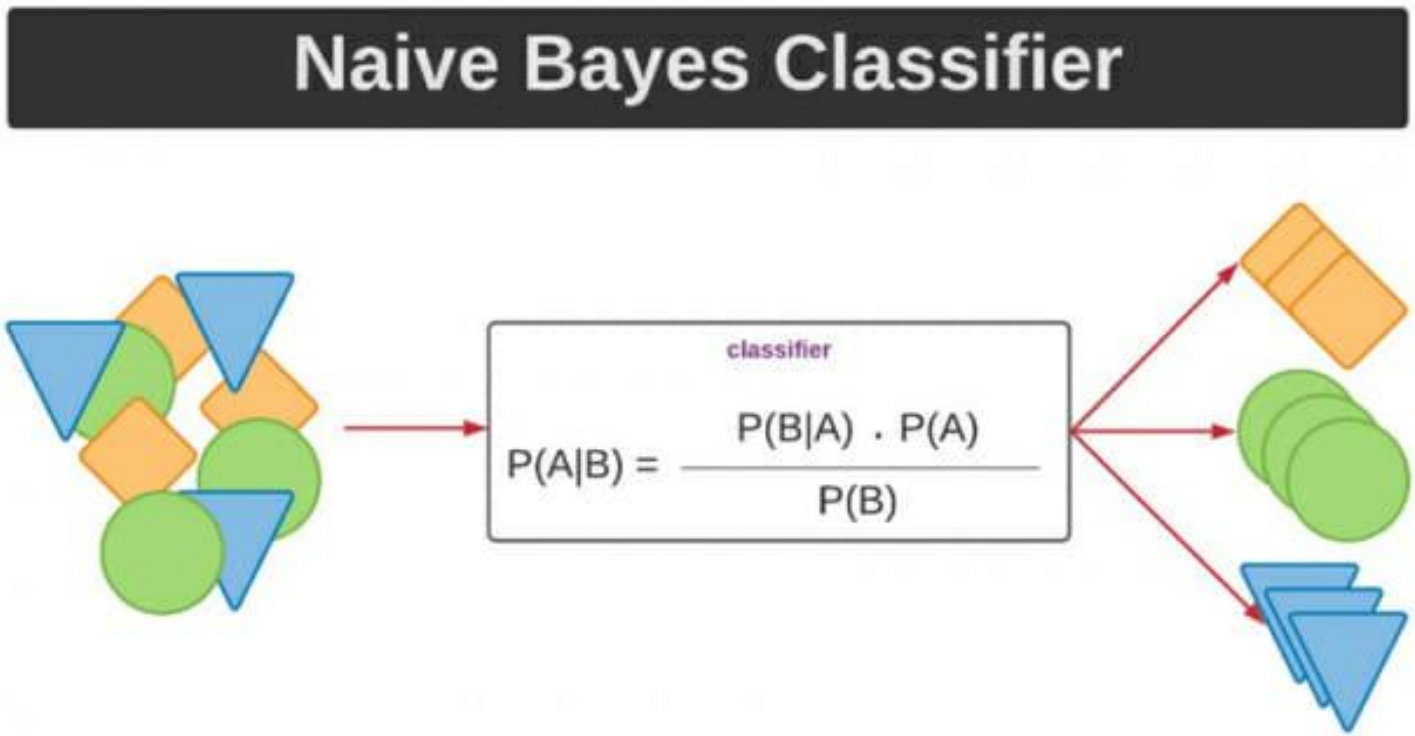
랜덤포레스트



Accuracy : 0.9121, Precision : 0.9199
Recall : 0.9809, F1 : 0.9619, Auc:0.8036

Chapter.3

나이브베이즈 분류기



복잡하게 섞여 있는 문제를 비슷한 성격을 가진 특성(feature)으로 분류

Chapter.3

하이퍼파라미터

- 나이브베이즈

params={'alpha' : [0.5, 1.0, 1.5, 2.0]. 'fit_prior' : [True, Fasle]}

-> 최적 하이퍼 파라미터

: {'alpha' : 1.0 , 'fit_prior' : True} ,

최적 모델 평균 성능 : 0.93285

```python
from sklearn.model_selection import GridSearchCV
from sklearn.naive_bayes import MultinomialNB
model = MultinomialNB()

params = { "alpha" : [0.5,1.0,1.5,2.0] , "fit_prior": [True, False] }


grid_clf = GridSearchCV(model, param_grid=params, cv=10)
grid_clf.fit(X3_train, y3_train)

print(f'최적의 하이퍼 파라미터 : {grid_clf.best_params_}')
print(f'최적의 모델 평균 성능 : {grid_clf.best_score_}')
```

✓ 1.3s

```
최적의 하이퍼 파라미터 : {'alpha': 1.0, 'fit_prior': True}
최적의 모델 평균 성능 : 0.9328580344280726
```

Chapter.3

다른 호텔 리뷰에 적용

2.0

Kyoungmi (대한민국)
커플/2인 여행객
스탠다드 더블룸
2023년 2월 | 2박

"무난함"

깔끔하고 위치도 훌륭했어요 다음에 또 가고 싶은 호텔입니다

작성일: 2023년 2월 26일

특가 상품 보기

긍정

```
input='깔끔하고 위치도 훌륭했어요 다음에 또 가고 싶은 호텔입니다'

def predict_str(input):
    input = re.sub(r'[^ㄱ-ㅎㅏ-ㅣ가-힣 ]','',input)
    tagger=Okt()
    morphs=tagger.morphs(input)
    morphs = [x for x in morphs if len(x) > 1]
    morphs = [x for x in morphs if x not in stoplist]
    morphs = [' '.join(morphs)]

    n=vectorizer2.transform(morphs)
    predict=model3.predict(n)
    print(predict)

predict_str(input)
```

✓ 0.1s

```
[1.]
```

Chapter.3

다른 호텔 리뷰에 적용

6.8 양호

sunok (대한민국)
유아/아동 동반 가족 여행객
스탠다드 트윈
2018년 2월 | 1박

"위생, 서비스에 조금만 신경쓰시면 더 좋을 것 같아요."

전망이 좋지 않아요. 이불이 좀 지저분 했어요. 체크인 할 때 여직원이 불친절해서 기분이 상했습니다. 대신 호텔과 몰이 연결되어 있고 식당가가 많아서 편리했어요.

작성일: 2018년 2월 4일

특가 상품 보기

부정

```
1  input='전망이 좋지 않아요. 이불이 좀 지저분 했어요. 체크인 할 때 여직원이 불친절해서 기분이 상했습니다. 대신 호텔과 몰이 연결되어 있고 식당가가 많아서 편리했어요.'
2
3  def predict_str(input):
4      input = re.sub(r'[^ㄱ-ㅎㅏ-ㅣ가-힣 ]','',input)
5      tagger=Okt()
6      morphs=tagger.morphs(input)
7      morphs = [x for x in morphs if len(x) > 1]
8      morphs = [x for x in morphs if x not in stoplist]
9      morphs = [' '.join(morphs)]
10
11     n=vectorizer2.transform(morphs)
12     predict=model3.predict(n)
13     print(predict)
14
15 predict_str(input)
```

0.1s

```
[0.]
```

Chapter.4

결과

- 평점 7점 기준 긍정/부정 나누기

- 나이브베이즈 모델로 평점 6점 미만 9점 이상 데이터 학습

- 머신러닝 결과로 다른 호텔 리뷰 적용하여 긍정인지 부정인지 확인

Chapter.5

결론

- 평점이 없는 새로운 리뷰를 긍정/ 부정 구분할 수 있게 됨

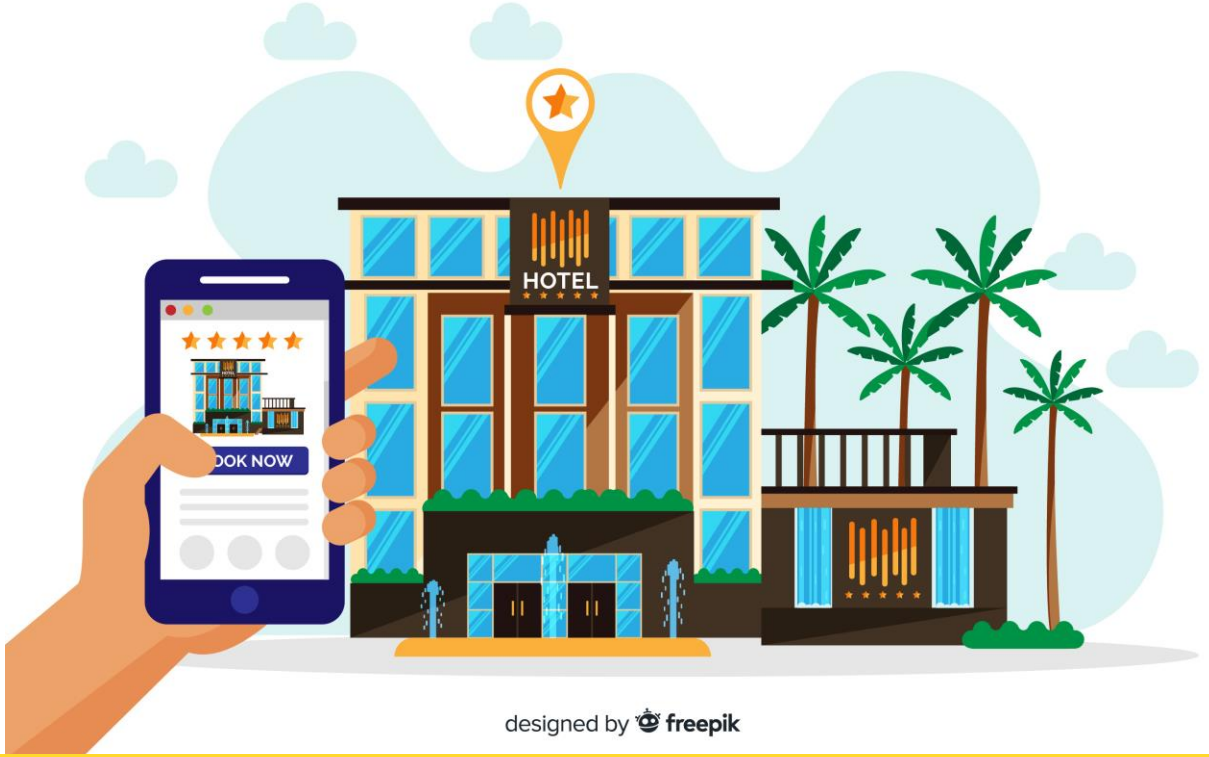
- 앞에서 언급한 평점의 모호한 기준 보완 가능

Chapter. 6

출처

- https://www.sap.com/korea/insights/what-is-machine-learning.html
- https://zdnet.co.kr/view/?no=20220725093548

감사합니다



designed by freepik