정보통신기획평가원 한경아카데미 과학기술정보통신부

실무 프로젝트 기반 빅데이터 전략 마에스트로 과정

BigData분석을 통한 막걸리 활성화 방안

맛걸리 프로젝트

김윤진 | 변수현 | 이병민 | 최유림

CONTENTS

막걸리 시장 현황
및 문제점

데이터 전처리

데이터 수집
데이터 정제
파생속성 생성
데이터셋 생성

데이터 분석

군집 분석(K-Medoids)
분류 분석(Gradient Boosting)

서비스 구현

1 막걸리 시장 현황 및 문제점

찹쌀·멥쌀·보리·밀가루 등을 쪄서 누룩과 물을 섞어 발효시킨 한국 고유의 술

1964 막걸리 쌀 사용 금지로 주질 저하

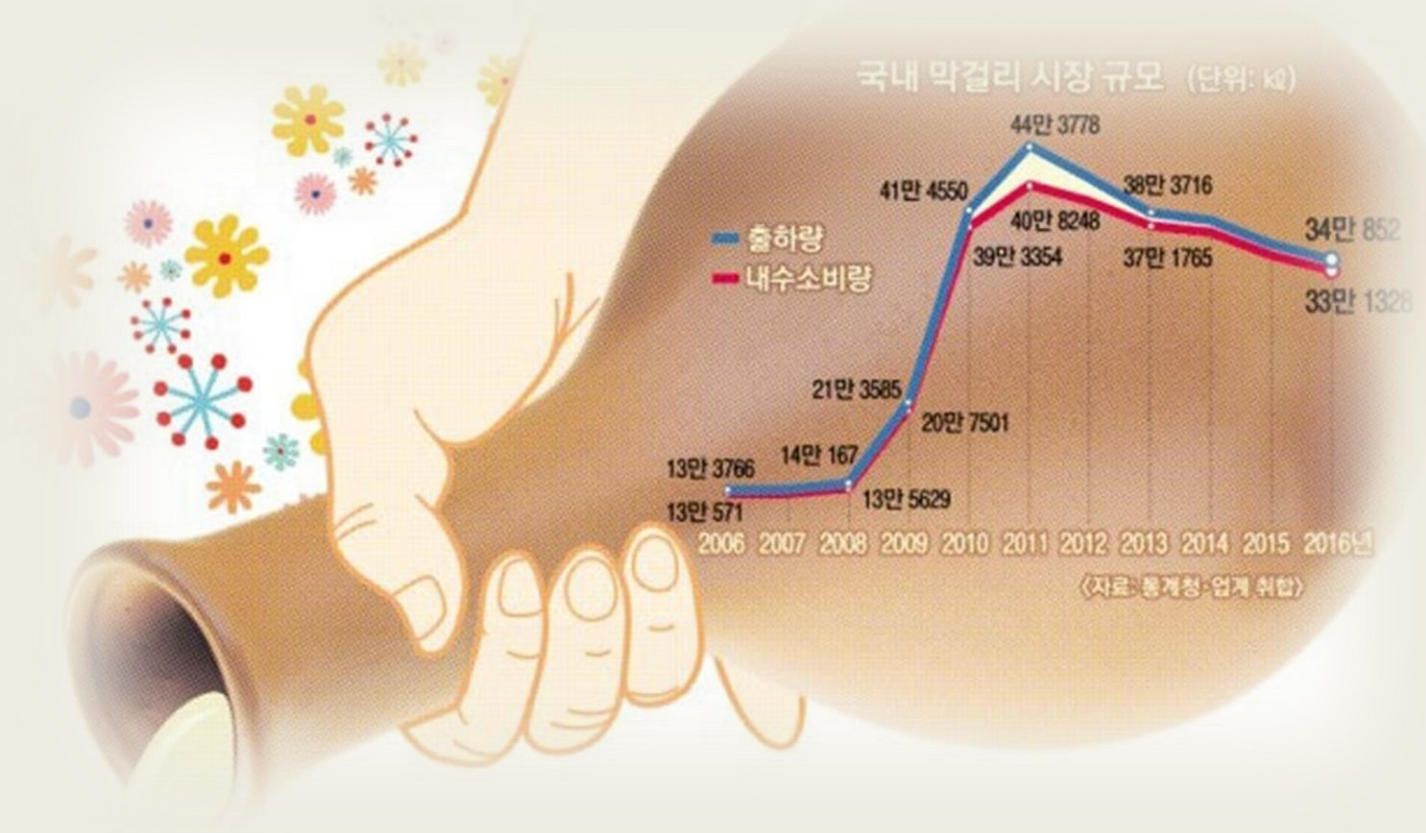
1971 쌀막걸리 재허가, 막걸리 주조방법 규격화

1972 쌀 가격 상승으로 쌀 막걸리 사라짐

2009 막걸리 열풍, 다양한 제품이 출시

출처 : 한국민족문화대백과(한국학중앙연구원)

1 막걸리 시장 현황 및 문제점

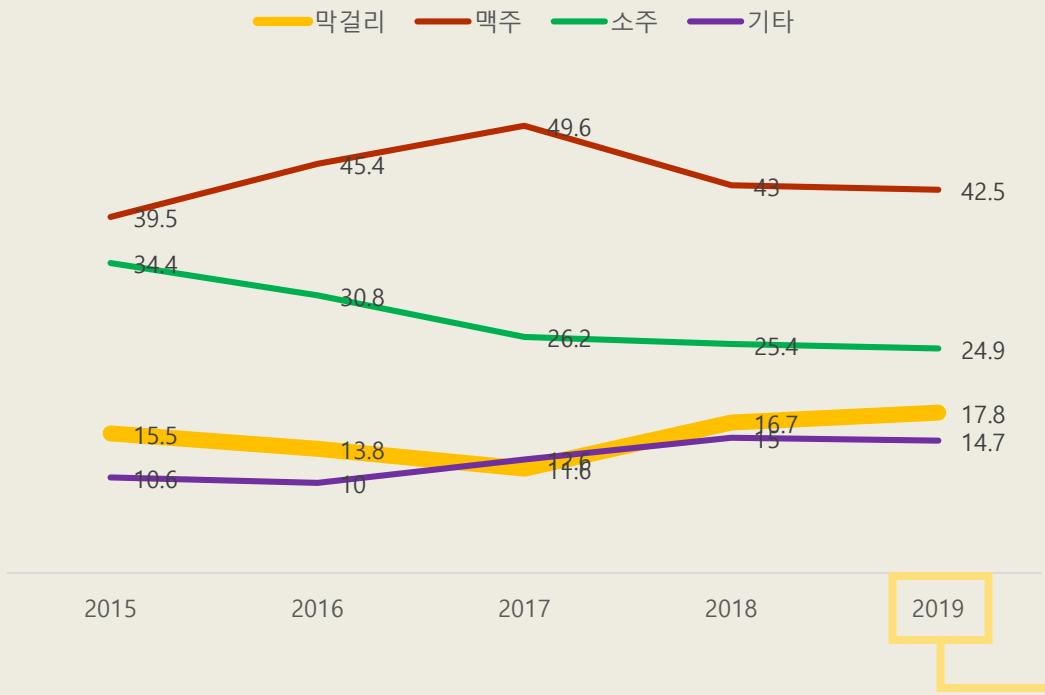


2011년 이후로 막걸리 출하량 및 내수소비량 지속적 감소

출처 : 한국민족문화대백과(한국학중앙연구원)

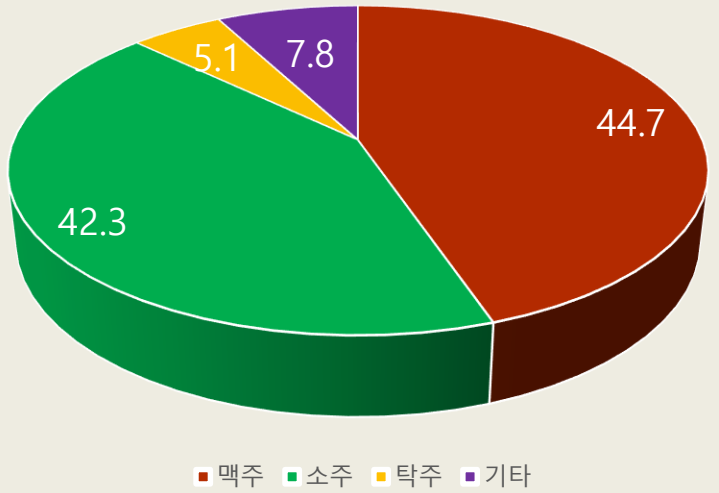
1 막걸리 시장 현황 및 문제점

연도별 선호 주종



막걸리보다 맥주와 소주를 각 2배, 1.5배 선호

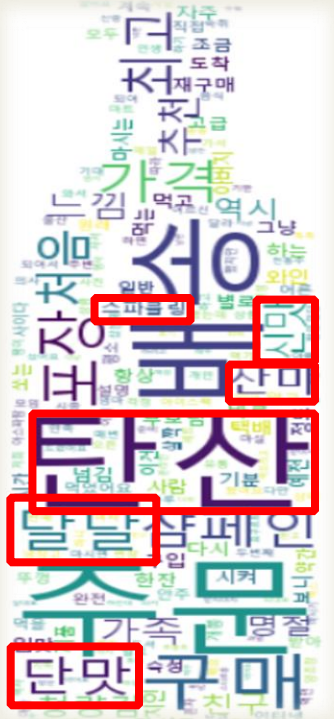
2019년 주종 별 국내 소비량



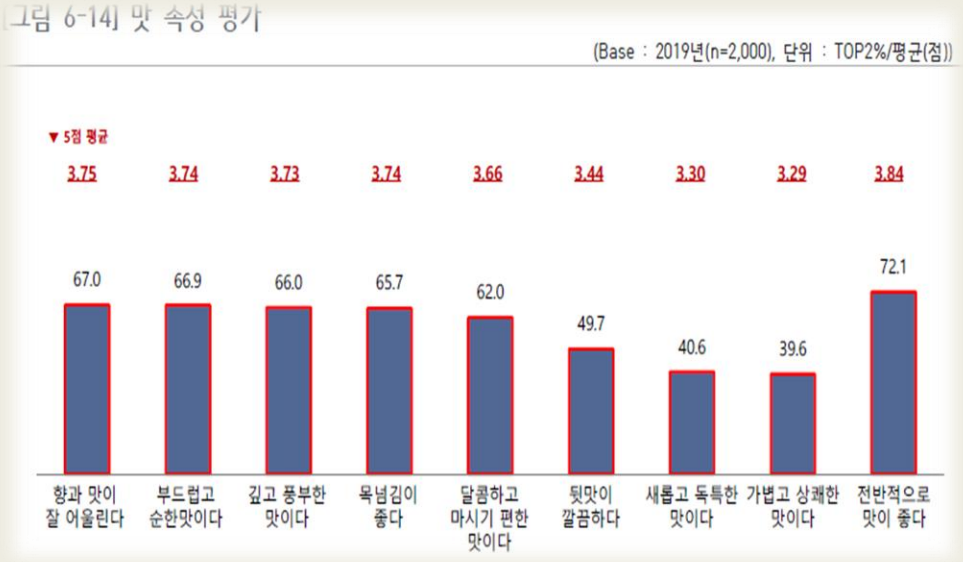
2019년 기준 막걸리는 주류 소비 중 5%만 차지

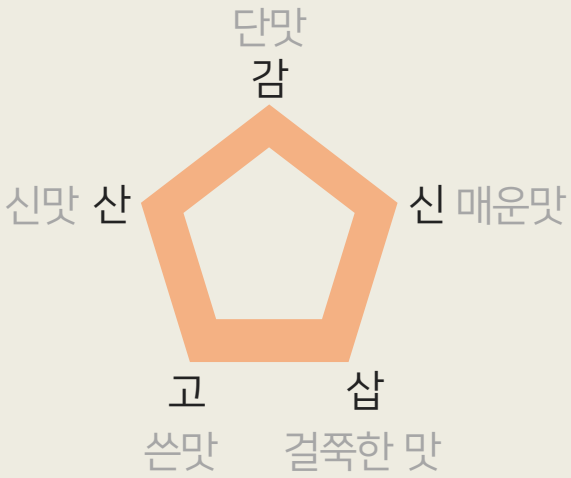
출처 : 2019 주류 시장 트렌드 보고서

1 막걸리 시장 현황 및 문제점



주류시장트렌드보고서 전통주 맛 속성 질문





맛걸리의 5가지 맛

막걸리 소믈리에 직업이 있을 정도로

막걸리는 다양한 맛을 가지고 있음

출처 : 네이버 쇼핑, 트래비매거진, : 2019 주류 시장 트렌드 보고서

1 막걸리 시장 현황 및 문제점

제조법/발효법에 따른 분류

색상/알코올 함량/당도에 따른 분류

출처 : 한국맥주문화진흥회, 한국주류산업협회

1 막걸리 시장 현황 및 문제점

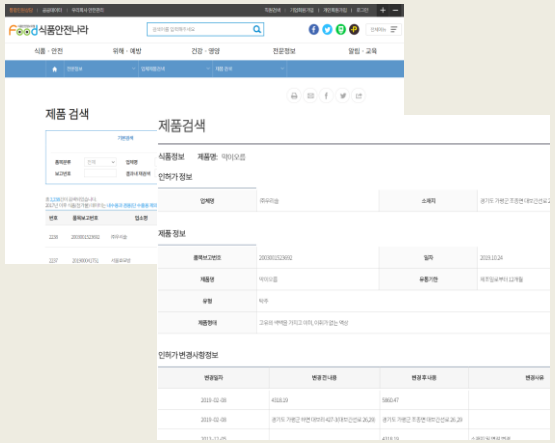
제조법/발효법에 따른 분류

색상/알코올 함량/당도에 따른 분류

“ 막걸리 세분화를 통하여
각 분류 별 특징을 정의 하는 것이 필요 ”

출처 : 한국맥주문화진흥회, 한국주류산업협회

2 데이터 전처리 | 데이터 수집



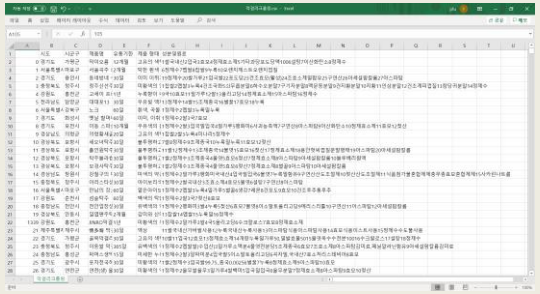
식품안전나라 막걸리 제품 확인

식품유형 : 탁주
소재지, 유통기한, 제품명, 제품형태,
성분 및 원료 등

데이터 크롤링

원본 데이터셋 생성

속성 : 소재지, 제품명, 유통기한,
제품형태, 성분 및 원료
행 : 2251개

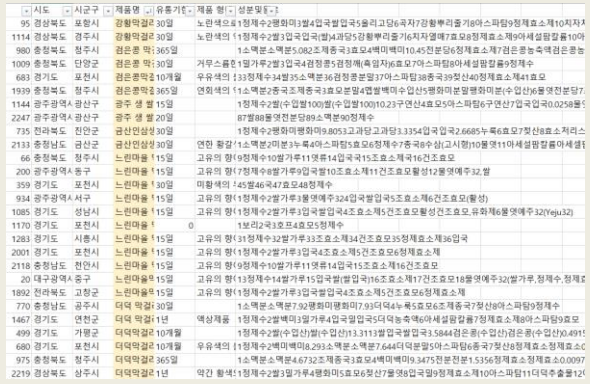


2 데이터 전처리 | 데이터 정제

중복 행 및 결측치 제거

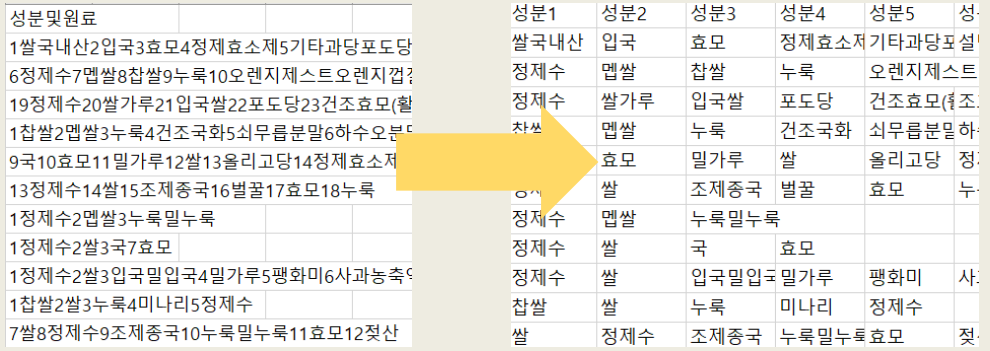
제품명, 유통기한, 소재지, 성분이 같은 행 제거

유통기한, 성분 등 결측치가 있는 행 제거

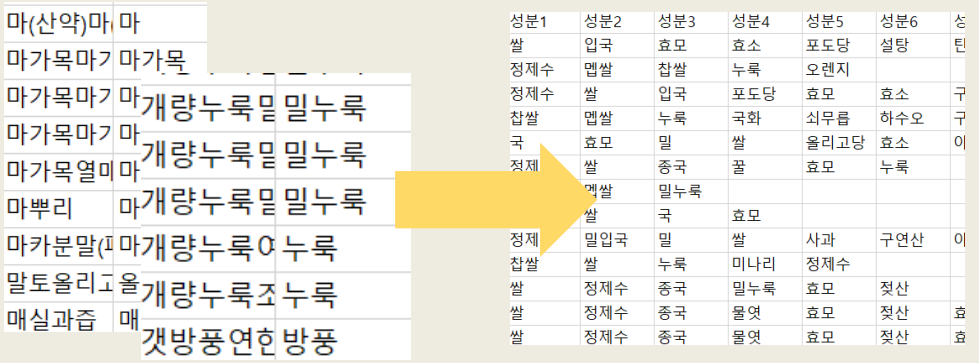


성분 컬럼 정리

1. 성분 컬럼 나누기



2. 성분명 정리하기



2 데이터 전처리 | 파생속성 생성

파생속성 생성

1. 성분

누룩 : 곡, 입국, 국, 종국

탄산 : 탄산, 일반

종류 : 막걸리, 동동주, 전통주, 청주

과일 : 오렌지,약재,율무...

주원료 : 쌀, 밀, 쌀&밀, 기타

2. 제품형태

색 : 백색, 미황색, 연미색, 유백색...

3. 유통기한

살균: 생, 살균

국	탄산	살균/생	종류(동동주	색	과일	단맛	주원료
입국	탄산	살균	막걸리	백색	0	설탕, 기타	쌀
누룩	일반	살균	막걸리	미황색	오렌지	0	쌀
입국	일반	생	막걸리	미황색	0	포도당, 아	쌀
누룩	일반	생	전통주	미황색	약재	0	쌀
국	일반	살균	막걸리	백색	0	올리고당,	쌀,밀
종국	일반	생	막걸리	우유색	0	벌꿀	쌀
누룩	일반	살균	막걸리	백색	0	0	쌀
국	일반	살균	막걸리	유백색	0	0	쌀
입국	탄산	살균	막걸리	우유색	0	아스파탐	밀
누룩	일반	생	전통주	미색	0	0	쌀
종국	일반	생	막걸리	백색	0	0	쌀
종국	일반	생	막걸리	백색	0	물엿, 아스	쌀
종국	일반	생	막걸리	백색	0	아스파탐,	쌀
종국	일반	생	막걸리	백색	0	물엿, 아스	쌀
입국	일반	생	막걸리	미색	0	물엿, 사카	밀
국	일반	생	막걸리	미색	0	물엿, 설탕,	쌀

제품 별 맛(단맛,신맛,고소한맛,드라이) 속성 생성

1. 성분 별 맛(단맛, 신맛, 고소한맛, 드라이) 확인
2. 제품 별 맛 컬럼 생성

성분	맛
개암가루	고소한맛
아스파탐	단맛
솔잎	드라이
밀	고소한맛
백미	
정제수	
효모	
스테비아	단맛
배추	고소한맛
복분자	신맛,
블루베리	단맛,

녹자	드라이
라벤더	드라이
옥수수	고소한맛
캐모마일꽃	드라이
후추	드라이
히비스커스	신맛
건포도	단맛
검은콩	고소한맛
검정깨	고소한맛
흑미	고소한맛

드라이	신맛	고소한맛	단맛	탄산
1	1	0	2	1
0	1	0	0	0
1	1	0	1	0
1	1	0	0	0
1	0	0	1	0
0	1	0	2	0
1	0	0	0	0
0	0	0	0	0
1	1	0	2	1
0	1	0	0	0
1	0	0	0	0
0	0	0	2	0
0	0	0	2	0
0	0	0	2	0
1	1	0	2	0
0	1	0	2	0

2 데이터 전처리 | 데이터셋 생성

데이터 범주화

누룩 : 곡 0/ 입국 1/ 종국 2/ 국 3
탄산 : 없음0/있음1
종류 : 막걸리 0/동동주1/전통주2/청주 3
주원료 : 쌀0/밀1/쌀밀2/기타3
살균 : 없음0/있음1
과일 : 없음0/있음1
단맛 : 단맛 없음()~아주 단맛(4)
신맛 : 없음0/있음1
드라이 : 없음0/있음1
고소한맛 : 없음0/있음1

데이터 셋 생성

1. 군집분석 : 주원료, 누룩, 살균, 탄산

2. 분류분석 : 단맛, 과일, 신맛, 드라이, 고소한맛

제품명	국	살균	주원료	탄산
막이오름	1	1	0	1
서울곡주오	0	1	0	0
동네방네수	1	0	0	0
청주신선주	0	0	0	0
고세이프리	3	1	2	0
대대포13	2	0	0	0
노크	0	1	0	0
옛날할머니	3	1	0	0
이동스파클	1	1	1	1
의령황새골	0	0	0	0
세오녀탁주	2	0	0	0
홀인원탁주	2	0	0	0
탁주블라썸	2	0	0	0
보경사탁주	2	0	0	0
강철구의미	1	0	1	0
아리스타싱	3	0	0	0

제품명	과일	드라이	신맛	고소한맛	단맛
막이오름	0	1	1	0	2
서울곡주오	1	0	1	0	0
동네방네수	0	1	1	0	1
청주신선주	1	1	1	0	0
고세이프리	0	1	0	0	1
대대포13	0	0	1	0	2
노크	0	1	0	0	0
옛날할머니	0	0	0	0	0
이동스파클	0	1	1	0	2
의령황새골	0	0	1	0	0
세오녀탁주	0	1	0	0	0
홀인원탁주	0	0	0	0	2
탁주블라썸	0	0	0	0	2
보경사탁주	0	0	0	0	2
강철구의미	0	1	1	0	2
아리스타싱	0	0	1	0	2

3 데이터 분석 | 군집 분석

K-Means

연속형 자료

샘플을 K개의 부분 집합으로 분리

현재의 분할이 군집이 되고
이 분할의 중심을 계산(평균)

각 객체를 가장 가까운 중심에 할당(거리)

다시 시작하여 더 이상
개체의 움직임이 없으면 마침

K-Medoids

범주형 자료

샘플을 K개의 부분 집합으로 분리

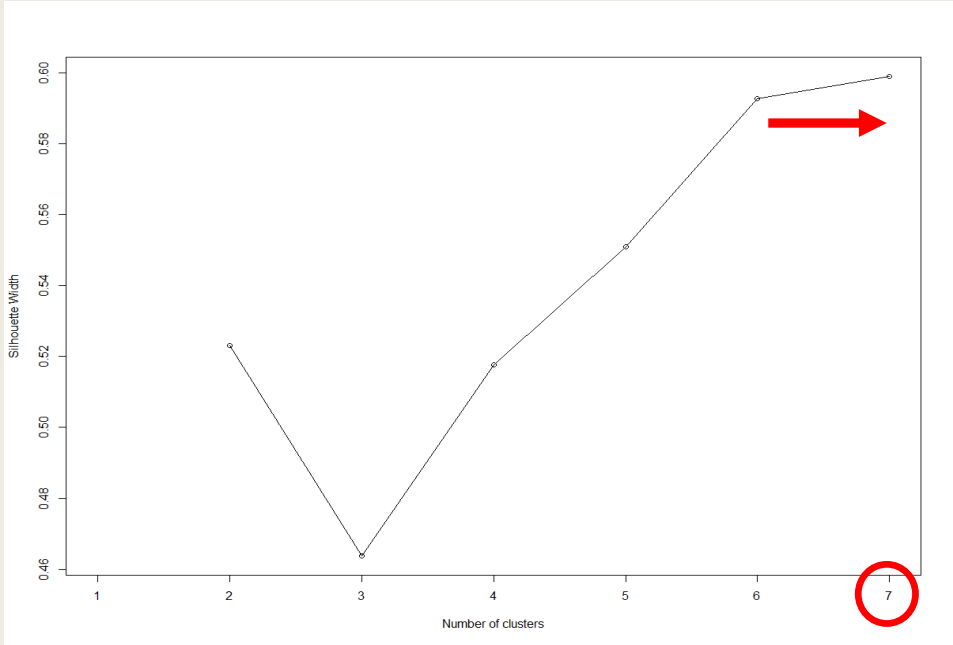
현재의 분할이 군집이 되고
이 분할의 중심을 계산(최빈값)

각 객체를 가장 가까운 중심에 할당(빈도)

다시 시작하여 더 이상
개체의 움직임이 없으면 마침

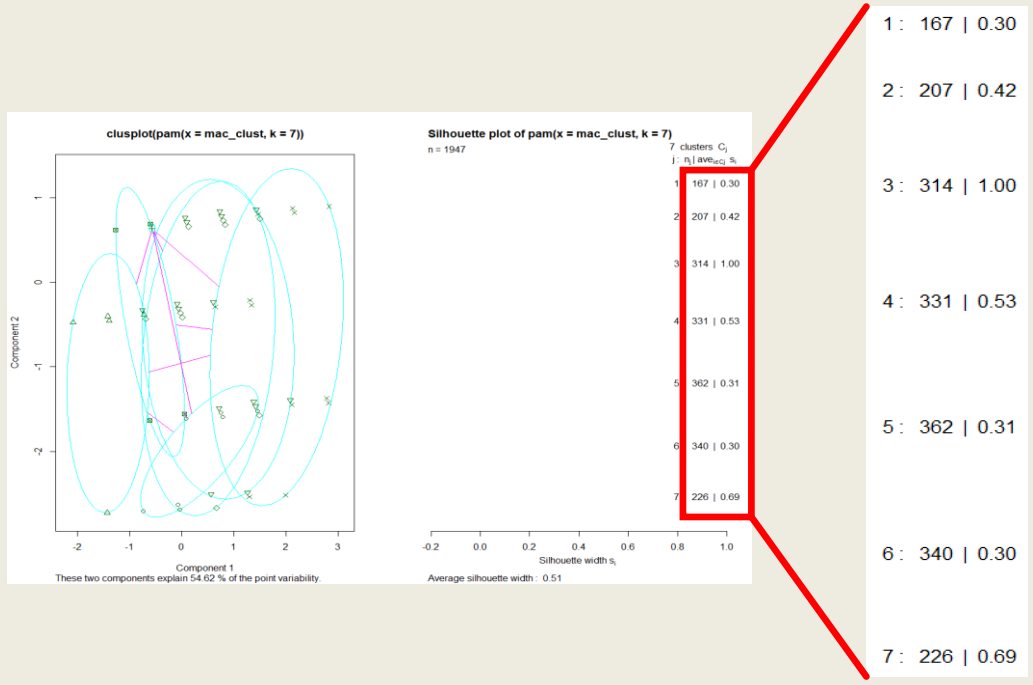
3 데이터 분석 | 군집 분석

1. 군집의 수 설정



dasiy함수를 통해 최적의 군집 개수 도출 : 7개
(Ward연결법 기준)

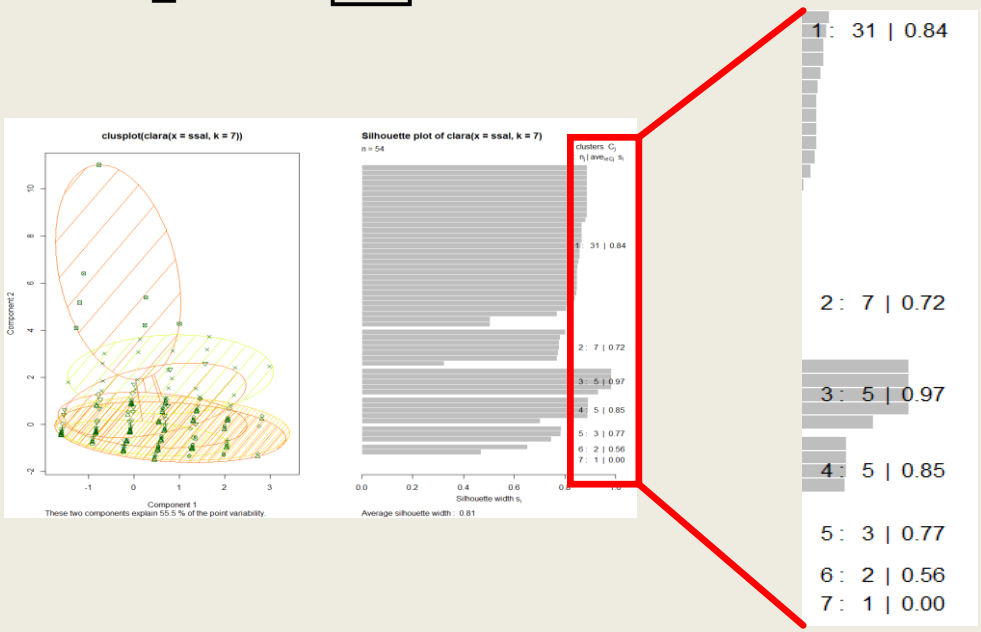
2. PAM함수



PAM함수의 경우 큰 데이터셋에 비효율적
군집 설명력이 좋지 않음

3 데이터 분석 | 군집 분석

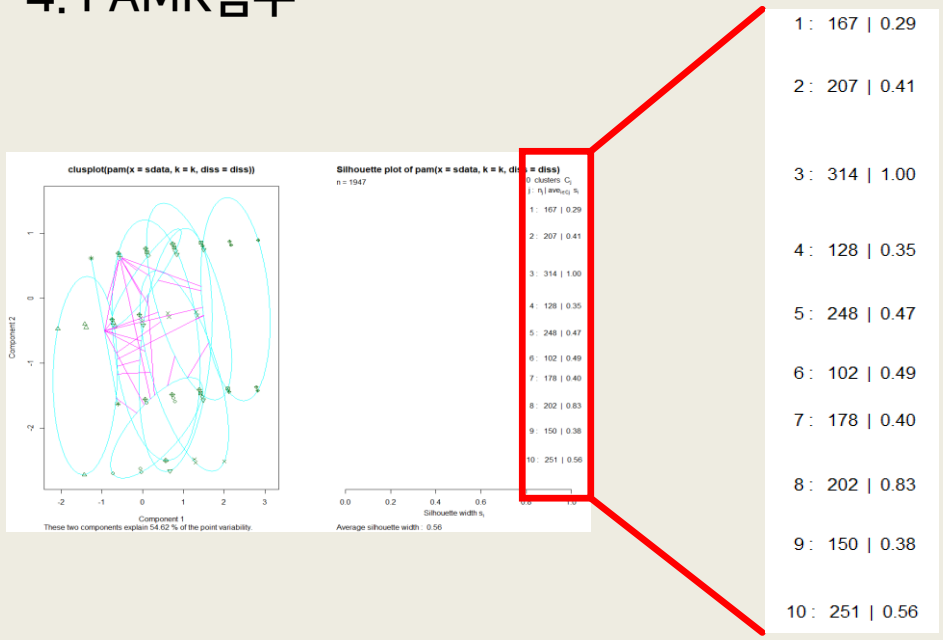
3. PAM_Clara



큰 데이터의 샘플을 가져와서 각 샘플에 PAM함수를 적용해
최선의 결과를 반환하는 CLARA 알고리즘

신뢰할 수 있는 군집 설명력

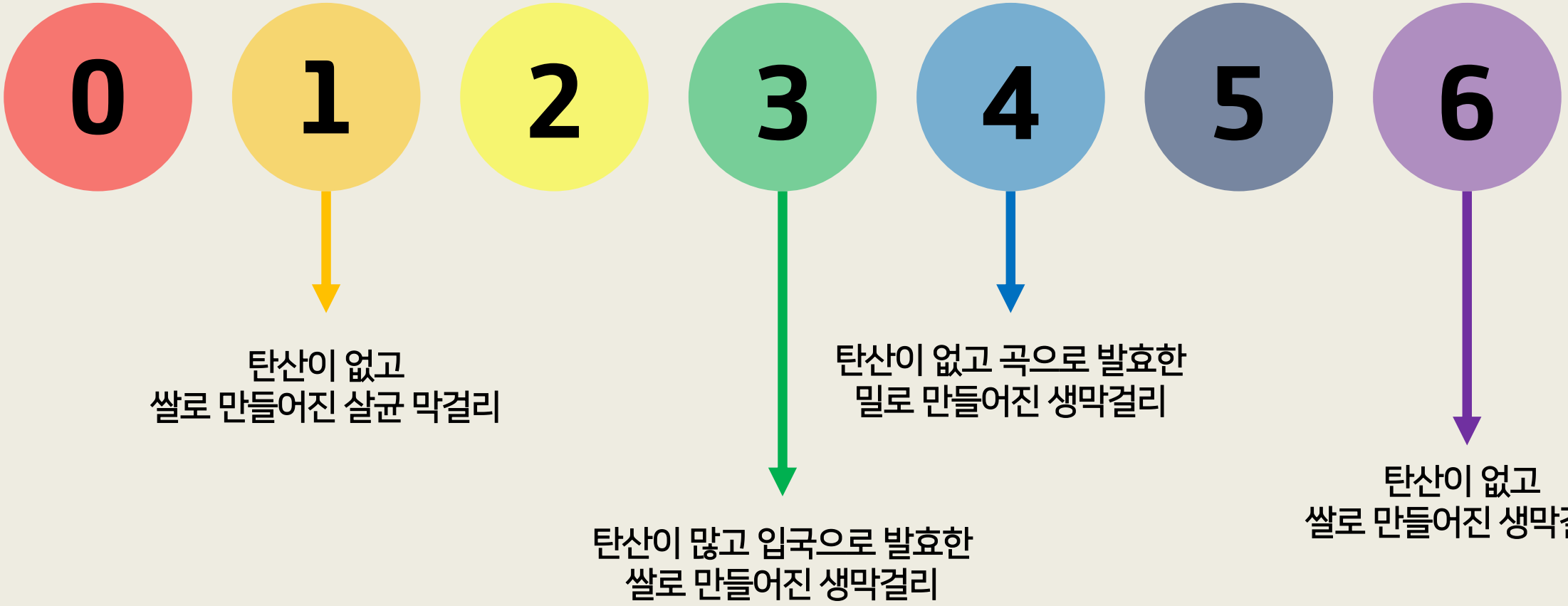
4. PAMK함수



군집의 수를 자동적으로 설정하여 군집분석을 실행

Clara보다 설명력이 낮음

3 데이터 분석 | 군집 분석


0
1
2
3
4
5
6
탄산이 없고
쌀로 만들어진 살균 막걸리
탄산이 많고 입국으로 발효한
쌀로 만들어진 생막걸리
탄산이 없고 곡으로 발효한
밀로 만들어진 생막걸리
탄산이 없고
쌀로 만들어진 생막걸리

3 데이터 분석 | 분류 분석

sparse_categorical_crossentropy

```
model = Sequential()
model.add(Dense(8, input_dim = 5, activation = 'softmax'))
model.add(Dense(8, activation = 'softmax'))
model.add(Dense(8, activation = 'softmax'))
model.add(Dense(7, activation = 'softmax'))

model.compile(loss = 'sparse_categorical_crossentropy', optimizer = 'adam', metrics=['accuracy'])

model.fit(x_train, y_train, epochs = 2000, verbose = 0)
```

```
<keras.callbacks.callbacks.History at 0x28bc2cc3ac8>
```

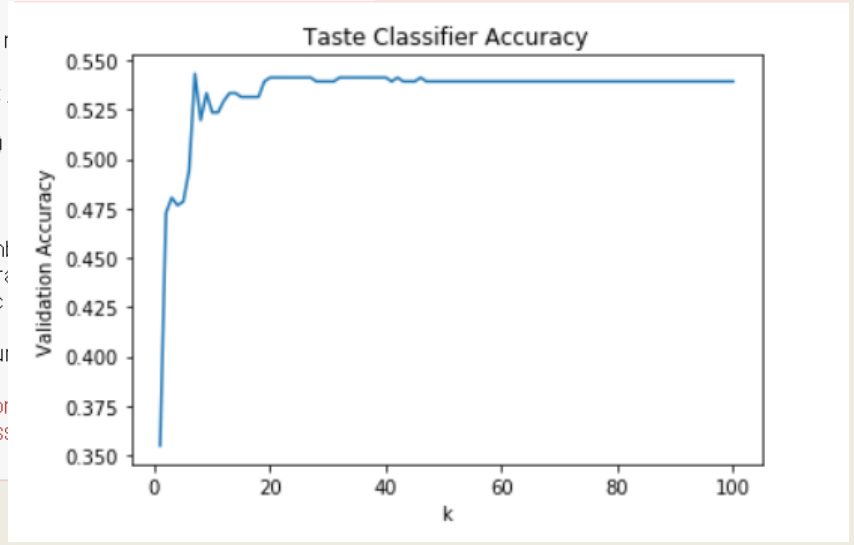
```
print("₩n 정확도 : %.2f" % (model.evaluate(x_test, y_test)[1]))
```

```
510/510 [==============================] - 0s 37us/step

 정확도 :  0.49
```

정확도 0.49

KNeighborsClassifier(KNN)

```
# KNN 분류 분석

classifier = KNeighborsClassifier(n_neighbors = 7)

classifier.fit(x_train

classifier.fit(x_test

k_list = range(1,101)
accuracies = []

for k in k_list:
  classifier = KNeighb
  classifier.fit(x_tra
  accuracies.append(c

plt.plot(k_list, accu
plt.xlabel("k")
plt.ylabel("Validation
plt.title("Taste Class
plt.show()
```



정확도 0.54

3 데이터 분석 | 분류 분석

Gradient Boosting ☑

```
# 사전가지치기와 학습률 설정하여 학습 실시
gbrt = GradientBoostingClassifier(random_state=0,
                                  max_depth=3,
                                  learning_rate=0.1,
                                  n_estimators=100,
                                  )

gbrt.fit(x_train, y_train)

print("훈련 세트 정확도 : {
print("테스트 세트 정확도 :
```

```
C:₩BigData₩anaconda3₩lib₩si
y was expected. Please char
  y = column_or_1d(y, warn=
```

```
훈련 세트 정확도 :    0.58
테스트 세트 정확도 :    0.54
```

```
# 사전가지치기와 학습률 설정하여 학습 실시
gbrt = GradientBoostingClassifier(random_state=0,
                                  max_depth=3,
                                  learning_rate=0.1,
                                  n_estimators=100,
                                  validation_fraction=0.1
                                  )

gbrt.fit(x_train, y_train)

print("훈련 세트 정확도 : {:.2f}".format(gbrt.score(x_train,y_train)))

print("테스트 세트 정확도 : {:.2f}".format(gbrt.score(x_test,y_test)))
```

```
C:₩BigData₩anaconda3₩lib₩site-packages₩sklearn₩ensemble₩_gb.py:1454: Da
y was expected. Please change the shape of y to (n_samples, ), for exar
  y = column_or_1d(y, warn=True)
```

```
훈련 세트 정확도 :    0.67
테스트 세트 정확도 :    0.65
```

정확도 0.65

동적프로그램 생성

```
드라이 한 것을 좋아하십니까?(YES:1/NO:0) : 0
신맛을 좋아하십니까?(YES:1/NO:0) : 0
고소한맛을 좋아하십니까?(YES:1/NO:0) : 1
단맛을 얼마나 좋아하십니까?(0~4) : 1

과일 막걸리를 좋아하십니까?(YES:1/NO:0) : 1
```

```
드라이 한 것을 좋아하십니까?(YES:1/NO:0) : 0
신맛을 좋아하십니까?(YES:1/NO:0) : 0
고소한맛을 좋아하십니까?(YES:1/NO:0) : 1
단맛을 얼마나 좋아하십니까?(0~4) : 1
과일 막걸리를 좋아하십니까?(YES:1/NO:0) : 1

[3]번 군집 막걸리 입니다

['이동마가리미주(중국수출전용)' '제주한라봉막걸리' '포천일동쌀막걸리(미국수출용)']막걸리를 추천합니다.
```

4 서비스 구현

식품유형	탁주
에탄올 함량	10%
원재료명 및 함량	쌀(국내산), 물, 국, 효모, 젖산(산도조절제), 밀 함유
내용량	750ml
업소명	(주)배혜정도가
소재지	경기도 화성시 정남면 서봉로 835
유통기한	병 어깨 표기일까지
보관방법	10℃이하 냉장 보관
품목보고번호	2013001702836

19세 미만 판매 금지
부정·불량식품 신고는 국번 없이 1399

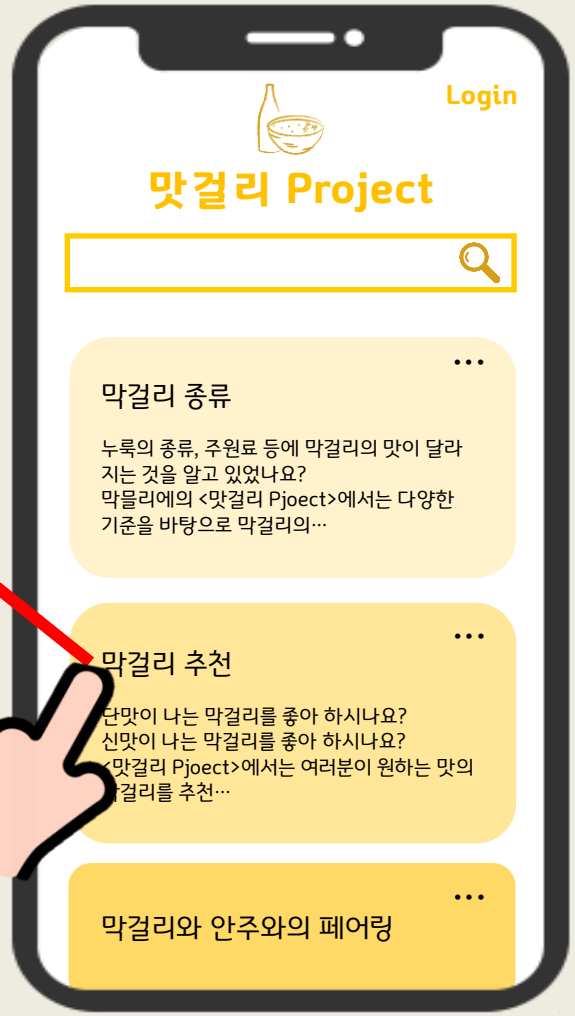
경고
지나친 음주는 뇌졸중, 기억력 손상이나 치매를 유발합니다.
임신 중 음주는 기형아 출생 위험을 높입니다.
제품 파손의 우려가 있으니
고온의 밀폐 공간에 놓아두지 마십시오.

A 막걸리(쌀)

성분표 하단에 막걸리의 군집을 표기해 막걸리 정보를 제공

4 서비스 구현


맛걸리 Project
막걸리 추천해 드림!
단맛
신맛
드라이
고소한맛
과일
원하는 맛을 클릭하면
막걸리의 군집과 막걸리를
추천해 드립니다!


Login
맛걸리 Project
막걸리 종류
누룩의 종류, 주원료 등에 막걸리의 맛이 달라
지는 것을 알고 있었나요?
막믈리에의 <맛걸리 Pjoect>에서는 다양한
기준을 바탕으로 막걸리의…
막걸리 추천
단맛이 나는 막걸리를 좋아 하시나요?
신맛이 나는 막걸리를 좋아 하시나요?
<맛걸리 Pjoect>에서는 여러분이 원하는 맛의
막걸리를 추천…
막걸리와 안주와의 페어링

Q&A