



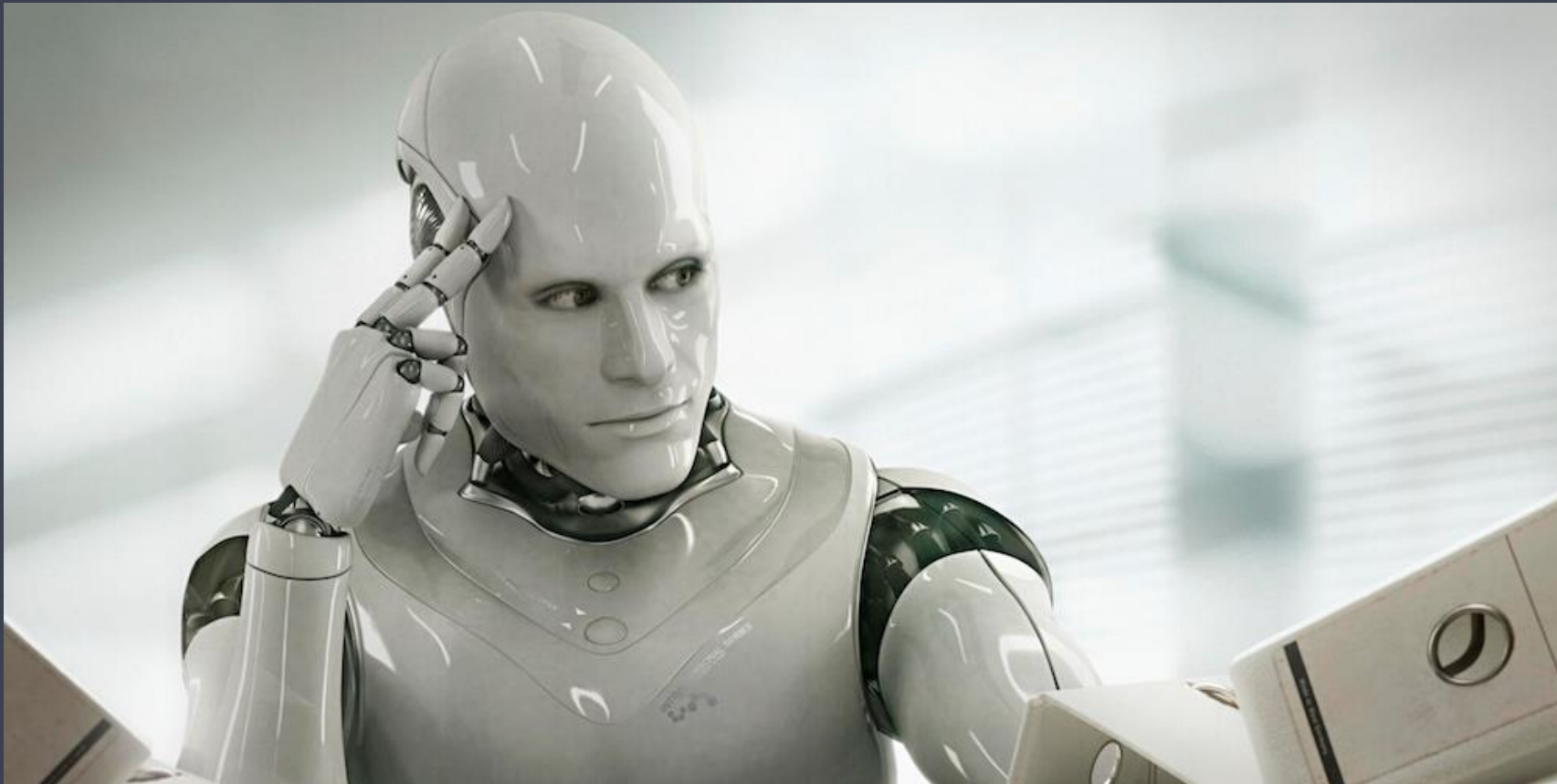
교재: 파이썬 머신러닝 완벽 가이드 - 위키북스 출판사 - 권철민저

제 1주	1.파이썬 기초	-변수와 임플렉 -조건문 반복문 -함수 -자료구조	실습
제 2주	2.파이썬 기반 머신러닝	-머신러닝 개념 -넵파이 -판다스	실습
제 3주	2.파이썬 기반 머신러닝	-matplotlib -seaborn -Aggregation 함수 ,GroupBy적용	실습
제 4주	3.사이킷런으로 시작하는 머신러닝	-사이킷런 소개와 특징 -붓꽃 품종 예측하기 -데이터 전처리 -사이킷런으로 수행하는 타이타닉 생존자 예측	실습
제 5주	4.평가	-정확도 -정밀도와 재현율 -F1스코어 -피마 엔디언 당뇨병 예측	실습
제 6주	5.분류	-결정트리 -양상불 학습 -랜덤포레스트 -분류 실습(캐글산탄테르 고객만족예측) -분류실습(캐글 신용카드 사기검출)	실습
제 7주	6.회귀	-회귀 개념 -비용 최소화 하기(경사하강법) -사이킷런을 이용한 보스턴 주택가격예측 -회귀실습(자전거 대여 수요예측)	실습
제 8주	중간고사	평가	실습
제 9주	6.회귀	-회귀실습(캐글 주택가격예측)	실습
제 10주	7.차원축소	-PCA -LDA -SVD -NMF	실습
제 11주	8.군집화	-k-평균알고리즘 개념 -군집평가 -군집화 실습(고객세그멘테이션) 고객세그멘테이션의 정의와 기법 데이터세트 로딩과 데이터 클렌징 RFM기반 데이터 가공 RFM 기반 고객 세그멘테이션	실습
제 12주	9.텍스트 분석	-텍스트 분석 이해 -텍스트 정규화 -BOW -텍스트분류 실습(10 뉴스그룹분류)	실습
제 13주	머신러닝 종합예제	-프로야구 선수 다음 해 연봉 예측 -비트코인 시세 예측	실습
제 14주	머신러닝 종합예제	-중고나라 휴대폰 거래가격 예측 -구매 데이터를 분석하여 상품 추천	실습
제 15주	기말고사	평가	평가

Machine Learning

Chapter 1 머신러닝 개요

- Machine Learning 개념을 이해 할 수 있다.
- Machine Learning의 종류 및 과정을 알 수 있다.



CHATTING ROBOT
SIMSIMI



TOUCH ME!

CHATTING ROBOT
SimSimi
SINCE 2002



ISMAKER

심심아 고민있어



심심이

저에게 말해보세요 다 들어드릴게요

여자친구랑 100일인데 뭐해줘야 좋아할까?



심심이

멋지게 헤어져

넌 재밌냐 사는데



싱싱이

재밌겠냐 하루종일 대답이나 하고
앉아있는데...



데이터 특성과 패턴을 학습하여,
미지의 데이터에 대한 결과를 예측하는 것

지도학습 (Supervised Learning)

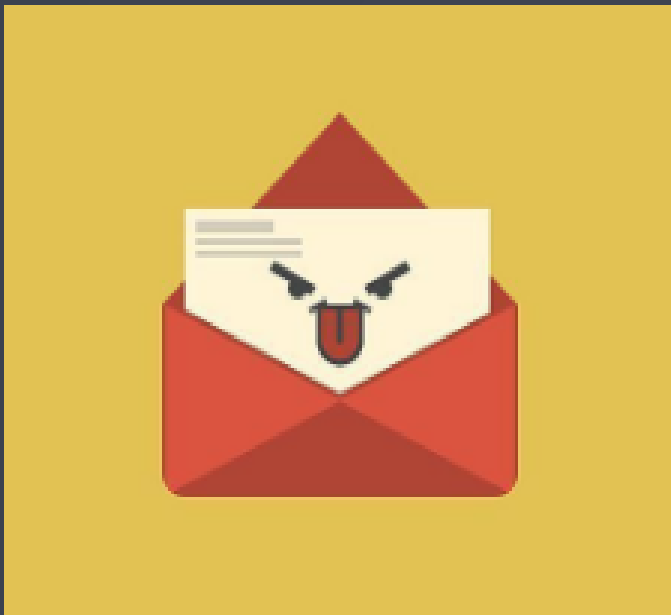
비지도학습 (Unsupervised Learning)

강화학습 (Reinforcement Learning)

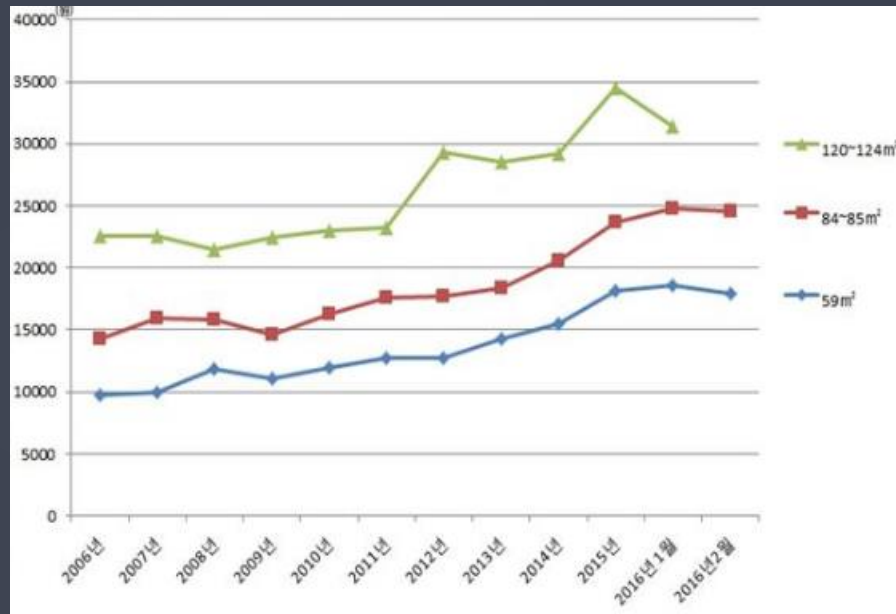
지도 학습 (Supervised Learning)

- 데이터에 대한 Label(명시적인 답)이 주어진 상태에서 컴퓨터를 학습시키는 방법.
- 분류(Classification)와 회귀(Regression)로 나뉘어진다.

지도 학습 (Supervised Learning)



스팸 메일 분류

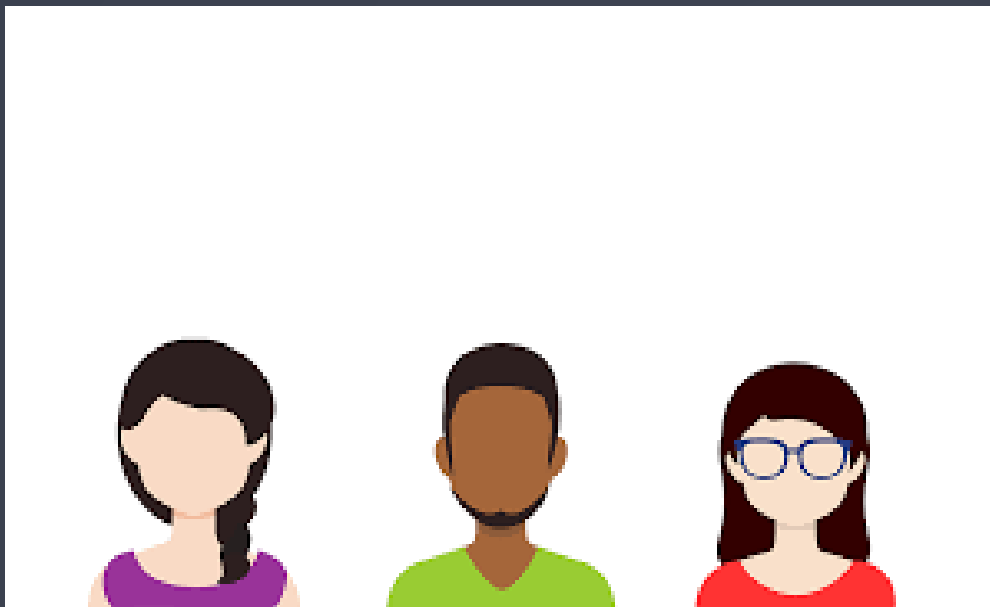


집 가격 예측

비지도 학습 (Unsupervised Learning)

- 데이터에 대한 Label(명시적인 답)이 없는 상태에서 컴퓨터를 학습시키는 방법.
- 데이터의 숨겨진 특징, 구조, 패턴을 파악하는데 사용.
- 데이터를 비슷한 특성끼리 묶는 군집(Clustering)과 차원 축소(Dimensionality Reduction)등이 있다.

비지도 학습 (Unsupervised Learning)

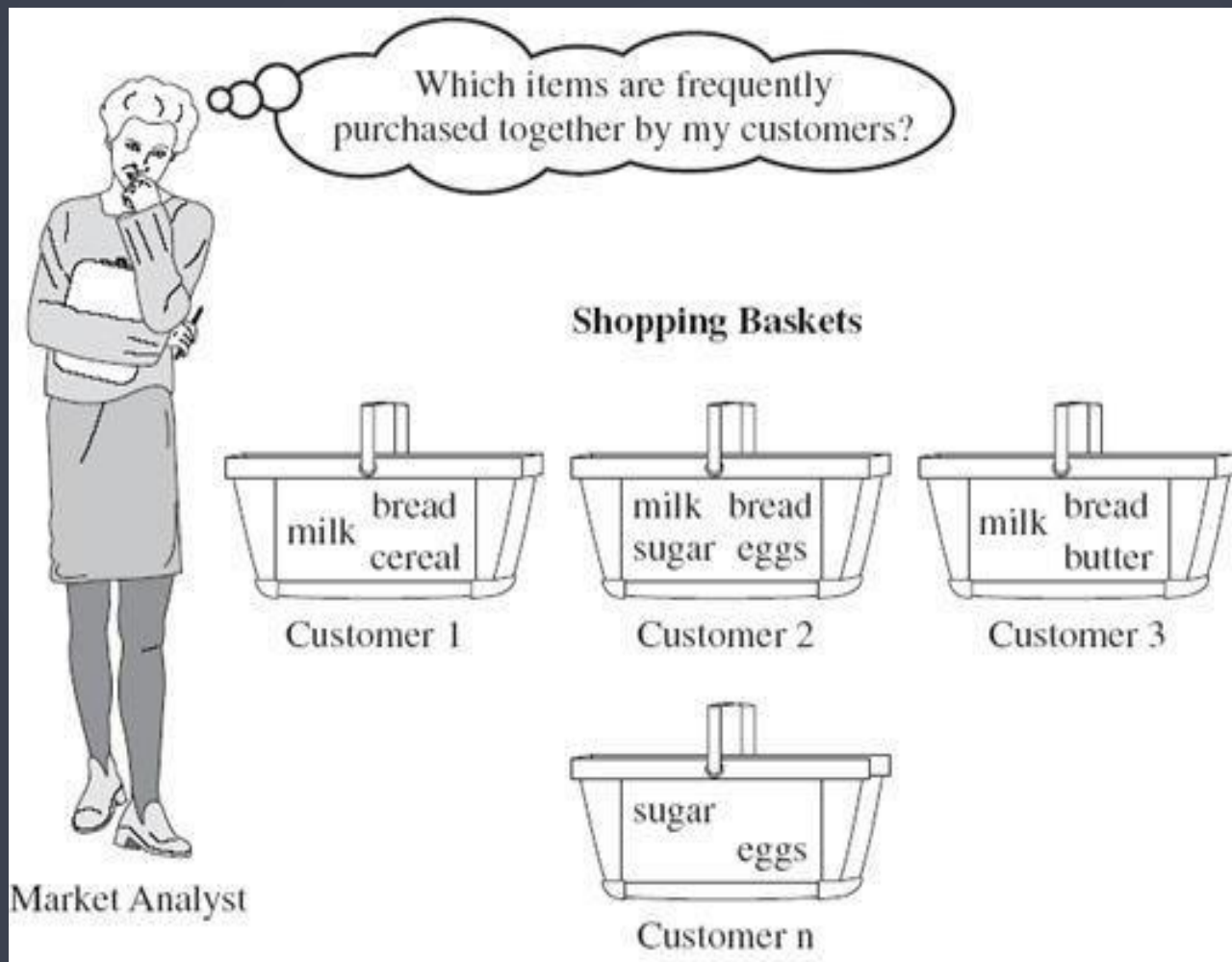


소비자 그룹 발견을 통한 마케팅



글 주제 구분

비지도 학습: 연관규칙 (Association Rules)



강화 학습 (Reinforcement Learning)

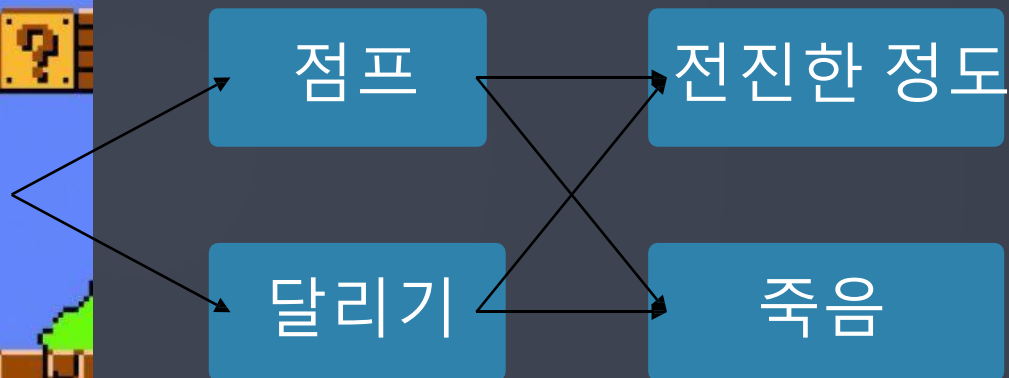
- 문제와 답을 주지 않고 목표와 보상만 제공하여 컴퓨터를 학습시키는 방법.
- 기계는 더 많은 보상을 얻을 수 있는 방향으로 행동을 학습.
- 주로 게임이나 로봇을 학습시키는데 많이 사용.

강화 학습 (Reinforcement Learning)

- 보상 시스템에 따라 최적의 액션 시퀀스(action sequence)를 결정하는 것



환경



액션

보상

02. 기계학습 응용사례

금융 분야



신용평가모형을 활용한 자동 대출



음성 기반 금융 거래



고객의 사회적 데이터(예: 소셜 미디어 데이터)를 분석하여 고객의 성향을 파악

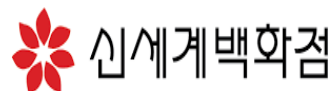


결제 정보와 주변 상권 정보를 학습하여 고객 동선과 점포 이용 성향을 분석



사람의 개입없이 투자하고 자산을 재구성하는 QV 글로벌 로보어드바이저 출시

소비/유통 분야



고객의 구매 기록, 성별, 나이, 지역 등의 데이터를 분석하여 맞춤형 광고 수행



음성인식 통역 소프트웨어를 탑재해 외국인과 소통할 수 있는 쇼핑 봇 활용 (동대문점)

ZARA 전 세계 매장의 판매 및 재고 데이터를 실시간 분석하는 '재고 최적 분배 시스템'을 개발



소비자의 소비 패턴 분석 기반의 '예측배송' 서비스를 도입



혼잡 시점과 혼잡 구역을 예측하고 배송지에 따른 최적 적재 경로를 안내

제조 분야

VOLVO

출고된 자동차에 데이터 수집 센서를 부착하여 비용 절감 및 불량 원인 감지

SK 하이닉스

반도체 생산 장비의 예지 정비

SAMSUNG
삼성전자

반도체 전체공정에서 수율에 영향을 미치는 인자 및 설비 발견


DAEDUCK

일일 수율 및 품질 현황, 월간/주간/일별 품질 현황 등의 다양한 품질 수율 분석


POSCO

불량 원인 파악 및 도금량 제어


공공 서비스 분야

서울특별시


30억 건의 통화량 데이터 분석을 통한 올빼미 버스 도입

기상청


기상 데이터 분석을 통한 효율적인 위험기상 예측

강남구
GANG NAM GU

주정차 민원 처리를 위한 '강남봇' 개발

관세청

불법 화물(예: 수입금지물품)을 통관 과정에서 선별

대한통운

대형폐기물 이미지 분류를 통한 과금 매칭 간편결제

1. Problem Identification(문제정의)
2. Data Collection(데이터 수집)
3. Data Preprocessing(데이터 전처리)
4. EDA(탐색적 데이터분석)
5. Model Selection(모델 선택)
6. Fit(학습)
7. Evaluation(평가)

1. Problem Identification(문제정의)

- Classification 분류
- Regression 회귀

2. Data Collection(데이터 수집)

- 공공데이터
(<https://www.data.go.kr/>)
- 웹크롤링 (뉴스, SNS, 블로그)
- Kaggle

3. Data Preprocessing(데이터 전처리)

- 결측치, 이상값 조정
- Encoding
Categorical Data를 수치 데이터로 변경
- Feature Engineering (특성공학)
단위 변환, 새로운 속성 추가

4. EDA(탐색적 데이터분석)

- 데이터를 관찰 후 전처리 전략 수립
→ 시각화(pandas, matplotlib, seaborn)
- 예측 모델에 넣을 Feature(특성) 결정

5. Model Selection(모델 선택)

- 목적에 맞는 적절한 모델 선택
- KNN, Decision Tree,
Linear Model, Ridge, Lasso ...
HyperParameter tuning
(하이퍼파라미터 조정)

6. Fit(학습)

- Train 데이터와 Test 데이터를
7:3 정도로 나눔

머신러닝(Machine Learning) 과정



7. Evaluation(평가)

- Accuracy(정확도)
- Mean squared error(평균제곱오차)

예시: 와인 품질 수준 파악

1. 문제 정의: 레드 와인의 기본 정보가 주어졌을 때, 와인의 품질 수준을 파악



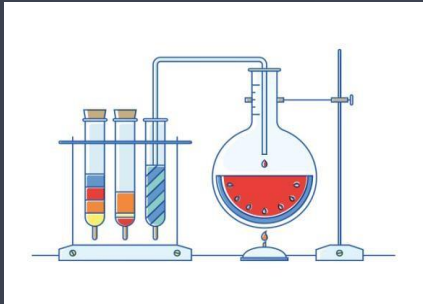
- 고정 산도
- 휘발성 산도
- 구연산
- 잔류 설탕
- 염화물
- 자유 이산화황
- 총 이산화황
- 밀도
- pH
- 황산염
- 알코올



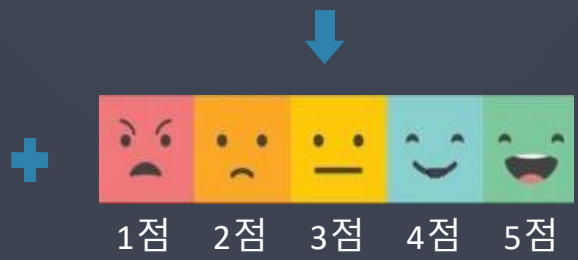
2. 데이터 수집: 물리 화학적인 실험을 통해 레드 와인의 기본 정보 데이터를 획득하고
,
소믈리에가 해당 와인의 품질을 입력 (기계가 소믈리에의 평가를 학습하는 것)

예시: 와인 품질 수준 파악

2. 데이터 수집: 물리 화학적인 실험을 통해 레드 와인의 기본 정보 데이터를 획득하고, 소믈리에가 해당 와인의 품질을 입력 (기계가 소믈리에의 평가를 학습하는 것)



- 고정 산도
- 휘발성 산도
- 구연산
- 잔류 설탕
- 염화물
- 자유 이산화황
- 총 이산화황
- 밀도
- pH
- 황산염
- 알코올

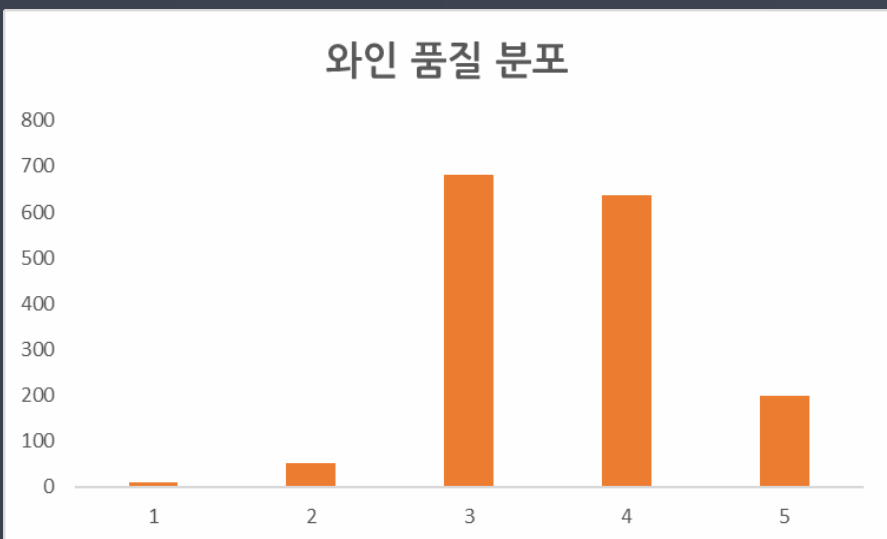


고정산도	휘발성산도	구연산	잔류설탕	염화물	자유이산화황	총이산화황	밀도	pH	황산염	알코올	품질수준
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	3
7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	3
7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	3
11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	4
7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	3
7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	3
7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	3
7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	5
7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	5

redwine_quality.csv

예시: 와인 품질 수준 파악

3. 데이터 탐색: 품질 수준의 분포 파악, 분포가 한쪽에 치우쳐져 있다면 정규화 작업



- 대부분 와인 품질이 3 혹은 4임을 확인
- 와인 품질이 1과 2인 경우가 너무 적고, 둘을 구분하는게 중요하지 않다고 판단
- 따라서 와인 품질 수준 변수를 1 (나쁨), 2 (보통), 3 (좋은)의 값을 갖도록 변경해야 함

4. 데이터 전처리: 와인 품질 수준 변수를 1 (나쁨), 2 (보통), 3 (좋은)의 값을 갖도록 변경

- 기존 와인 품질: 1 & 2 \Rightarrow 1
- 기존 와인 품질: 3 & 4 \Rightarrow 2
- 기존 와인 품질: 5 \Rightarrow 3