1. DM 과정

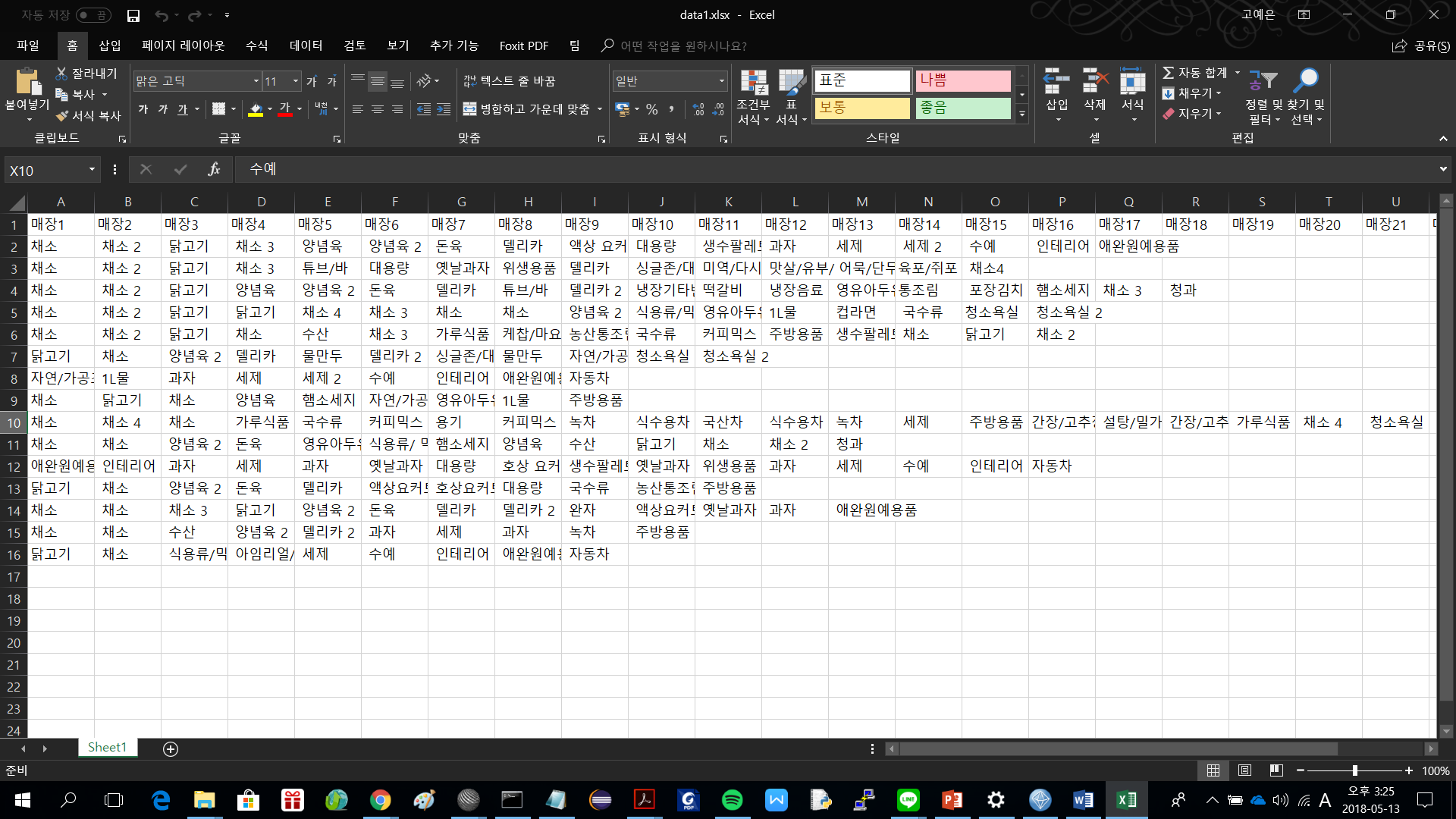
데이터(target data set) 수집🡪 데이터 다듬기 🡪 모델 개발&가설 설정🡪 DM 알고리즘 선택 🡪 결과 해석 & 시각화 🡪 검사&검증 🡪 관리

1. DM 작업(task)과 그 작업을 위해 사용해야 하는 알고리즘

* Classification: 많은 데이터들 중에 특정 특성(attribute)에 해당하는 값을 찾는 것
  + 사용 알고리즘: Decision tree, Memory based reasoning, K-nearest Neighbor
* Estimation: 데이터들의 기존 특성들을 바탕으로 새로운 값을 추출(예: 나이, 성별, 혈압 🡪 잔여수명). 미래 값을 추출하는 것이 Prediction
  + 사용 알고리즘: Neural Network
* Association: 데이터들 중에서, (항상) 같이 존재하는 값이 무엇인지 알아낸다.
  + 사용 알고리즘: Market Basket Analysis(규칙이 얼마나 좋은 지 판단)
  + Market Basket Analysis의 예:
    - Support: 고객이 오렌지를 구매한 후에 우유를 구매한 경우가 전체 고객 중 몇 %
    - Confidence: 오렌지를 구매한 고객 중 오렌지🡪우유를 포함하는 고객은 몇 %
    - Improvement: confidence가 높은 규칙이 좋은 규칙인지 판별
* Clustering: classification이 기존 정의된 카테고리가 있다면, clustering은 데이터들을 분석해서 새로운 카테고리를 만드는 것이다. (예: 증상들🡪질병, 고객 정보🡪판매할 물건). 예외 값들을 배제할 때 등에도 사용된다.
* 사용 알고리즘: K-Means 🡪 변주: Agglomerative, Weighted Distance, Gaussian mixture Models
* K-Means 사용 예:

1. 사용하는 도구: IBM SPSS
2. SPSS 에서 적용

* 논문(슈퍼마켓 고객의 동선분석)의 자료를 신경망(중에서 코호넨)을 이용해 분석
  + source



* + 결과: 14번째로 간 매장이 통조림, 16번째로 간 매장이 간장/고추장 코너일 때 연관성이 높다.

