제3장. 데이터 마트

1. 변수

1. **요약변수**

- 수집된 정보를 _ _ _ _ _에 맞게 _ _ _ _ _하는 변수
- 가장 기본적인 변수로 _ _ _ _ _ _ _ 금액, 횟수, _ _ _ _ _ 여부 등
- 많은 모델이 _ _ _ _ _으로 사용할 수 있어 _ _ _ _ _ _성이 높음
- 위클리 쇼퍼, 상품별 구매 순거, 단어 빈도, 트랜드 변수, 결측값 이상값 처리, 연속형 변수의 구간화

2. **파생변수**

- 사용자가 _ _ _ _ _ _ _을 만족하거나 _ _ _ _ _함수에 의해 값을 만들어 _ _ _ _ _를 부여하는 변수
- 매우 _ _ _ _ _적일 수 있으므로 _ _ _ _ _ _ _ _ _ _ _ _ _을 갖출 필요 있음
- 근무시간 구매지수, 주 구매 매장 변수, 시즌 선호 고객 변수, 라이프스타일 변수, 최적 통화시간 변수

2. 패키지

1. **Reshape 패키지**

- _ _ _ _ _ _ : 데이터를 _ _ _ _ _구조로 녹이는 함수
- _ _ _ _ _ _ : 새로운 구조로 데이터를 만드는 함수

2. **sqldf 패키지**

- R에서 _ _ _ _ _명령어를 사용가능하게 해주는 패키지
- SAS의 _ _ _ _ _ _ _ 과 같은 기능
- _ _ _ _ _ _ _ _ = _ _ _ _ _("select * from [df] limit 6")-
- _ _ _ _ _ _ _ _ = _ _ _ _ _("select * from [df] where [col] in ('BF', 'HF')")
- _ _ _ _ _ _ _ _ = _ _ _ _ _("select * from [df1], [df2]")

3. **plyr 패키지**

- _ _ _ _ _ 함수를 기반으로 데이터와 _ _ _ _ _변수를 동시에 배열로 치환
- _ _ _ _ _ _ _- _ _ _ _ _ _ _-_ _ _ _ _ _ 방식을 데이터를 분리, 처리, 결합
- 필수적인 데이터 처리 기능 제공

4. **Data Table 패키지**

- R에서 가장 많이 사용하는 _ _ _ _ _ _ _ _ _ _ _ 패키지 중 하나
- 대용량 데이터의 탐색, 연산, 병합에 유용
- 기존 _ _ _ _ _ _ _ _ 방식보다 월등히 빠른 속도
- 특정 칼럼을 _ _ _ _ _ 값으로 색인을 지정
- 빠른 _ _ _ _ _ _과 _ _ _ _ _ _, 짧은 문장 지원에서 유용

3. 데이터 가공 및 데이터 관리

1. **변수의 구간과**

- _ _ _ _ 모형 또는 고객 _ _ _ _ _ _ 등의 시스템으로 모형을 적용
- 각 변수들을 _ _ _ _ 하여 점수를 적용하는 방식
- _ _ _ _ _ _ : 연속형 변수를 범주형 변수로 변환하기 위해 _ _ _ _ 이하의 구간에 동일한 수의 데이터를 할당하여 구간을 _ _ _ _ 하는 방법
- _ _ _ _ _ _ _ _ : 모형을 통해 연속형 변수를 범주형 변수로 변환하는 방법

2. **결측값 처리**

- 변수에 데이터가 비어있는 경우 : _ _ _ _ , 999999, Unkown, _ _ _ _ _ _ _ _ 등
- _ _ _ _ _ _ (_ _ _ _ _ _ _ _ _ _ _)
  - Completes Analysis : 결측값의 레코드 삭제
  - _ _ _ _ _ _ _ _ _ _ _ : 관측 및 실험으로 얻은 _ _ _ _ 으로 대치
    - 비조건부 _ _ _ _ _ _ _ _ : 관측 데이터의 _ _ _ _ 으로 대치
    - 조건부 _ _ _ _ _ _ _ _ : _ _ _ _ 분석을 통해 데이터를 대치
  - _ _ _ _ _ _ _ _ _ _ _ : _ _ _ _ _ _ _ _ 에서 추정량의 표준오차의 _ _ _ _ _ _ _ _ _ _ 문제를 보완한 방법
    - Hot-Deck, Nearest Neighborhood
- _ _ _ _ _ _ _ _ (_ _ _ _ _ _ _ _ _ _ _) : _ _ _ _ _ _ _ _ 을 n번 실행하여 m개의 가상적 자료를 만들어 대치

3. **이상값 처리**

- bad data : 잘못 입력된 값이나 _ _ _ _ _ _ 에 부합되지 않는 값인 경우
- _ _ _ _ : 의도하지 않은 현상으로 입력된 값, 의도된 극단값
  - 평균으로 부터 _ _ _ _ _ _ _ _ 떨어진 값
  - 기하평균보다 _ _ _ _ _ _ _ _ 이상 떨어진 값
  - _ _ _ _ _ _ 와 _ _ _ _ _ _ 값에서 범위보다 2.5배 떨어진 값
  - _ _ _ _ (_ _ _ _ _ _) : 레코드 삭제
  - _ _ _ _ (_ _ _ _ _ _) : 상한 또는 하한값으로 조정