[3과목 문제풀이]
*데이터 분석 단계
요건정의-모델링-검증 및 데스트-적용단계
ㄴ요건정의
- 데이터 분석 업무의 배경, 주요 이유, 기대효과 파악, 분석수행 타당성 확인, 분석기법,절차 식별 및 구성, 분석방법론 구축
(오답: 분석대상 데이터 획득. 모델링 단계서 진행)
- 분석요건
 정의: 문제 해결 시 투자수익(ROI)로 증명할 수 있어야 한다.
- 분석요건을 구체적으로 도출, 선별, 결정하고 분석 과정을 설계 구체적인 내용을 실무 담당자와 협의하는 업무이다.
-이슈리스트 작성, 핵심 이슈 정의, 해결방안 정의 등이 주요 수행 업무이다.
(오답주의: 비즈니스 이슈 등 특정 요소에 많은 시간을 할애하면 안됨- 전체 프로세스에 저해)

*히스토그램을 잘 드러내는 통계량 : 사분위수 범위
- 사분위수 범위 : 산포의 측도
(히스토그램이 분포가 명확히 구분되는 경우)
- 평균,중앙값 : 위치 측도

* 측정수준과 -- 사용 가능한 기술통계
순서척도 -- 범위
구간척도 -- 최빈값
비율척도 -- 표준편차
(오답: 명목척도 - 중앙값x 성별,출신지 변수 존재)

*다중회귀분석 : 독립변수의 수가 너무 많아질 경우
- 설명력 증가가 현저히 감소
-추정치의 표준오차 증가
-회귀식의 적합도,타당도 감소
(오답: 종속변수에 대한 독립변수의 상대적 영향력 비교 곤란)

*의사결정나무
〈모형 특징〉 -비모수적 방법, 설명 용이, 계산 단순/빠름
(오답: 비정상적+잡은 데이터 민감)

〈불순도 impurity 측도〉 -지니계수, 엔트로피계수, 분류오류율
(오답: 감마계수---〉 연관성측도)

*지도학습 분석방법 :
- 의사결정나무, 베이지안 분류, 신경망 분석
(비지도학습--〉 k-평균 군집)

* k-means 군집분석
- k개의 평균을 중심으로 군(cluster)을 이루는 명령어
- 관측치, k개의 중심 간 거리 계산, 군을 이루는 명령어
- 군(cluster)의 개수는 사용자가 미리 결정한다.
- 계층적 구조를 보여주지는 않는다.
ㄴ 계층적 구조를 보여주는 분석 : hierarchical clustering

* 편상관(partial correlation)
- 변수x(연속형)와 변수y(연속형) 사이의 연관성을 살펴보고자 할 때, 제 3의 변수z(연속형)가 x와 y에 연관되어 있다고 가정 z에 조건화 하여 x와 y간 상관관계를 산출해야 할 때의 상관계수를 의미.

* 다차원척도법(Multidimensional Scailing)
- 여러 대상 간의 관계에 대한 수치적 자료를 이용해, 유사성에 대한 측정치를 상대적 거리로 시각화 하는 방법

* 연관성 분석 측정지표
- 아이템이 많아지면 어떤 연관규칙이 유의미한지 측정할 수 있는 평가지표가 필요
- 지지도(Support) : 전체 거래항목 중 상품 A와 상품 B를 동시에 포함하여 거래하는 비율
§ P(A ∩ B) : A와 B가 동시에 포함된 거래 수 / 전체 거래 수

- 신뢰도(Confidence) :
상품 A를 포함하는 거래 중 A와 B가 동시에 거래되는 비율
§ P(A ∩ B) / P(A) : A와 B가 동시에 포함된 거래 수 / A가 포함된 거래 수

- 향상도(Lift) :
상품 A가 주어지지 않았을 때 B의 확률 대비 A가 주어졌을 때 B의 확률 증가 비율
§ P(A ∩ B) / P(A) × P (B) = P(B | A) / P(B)
: A와 B가 동시에 일어난 횟수 / A와 B가 독립사건일 때, A와 B가 동시에 일어날 확률
§ A와 B 사이에 아무런 상호관계가 없으면 향상도는 1
§ 향상도가 1보다 높을 수록 연관성이 높다.
§ 즉, 향상도가 1보다 크면 B를 구매할 확률보다 A를 구매한 후 B를 구매할 확률이 더 높다는 의미

*감성분석(Sentiment Analysis) - Opinion mining
- 문장, 브랜드 평찬 특정 등 긍정/부정의 빈도 및 추이 확인이 가능하며, 각 문장이 긍정인지 부정인지를 주체에 따라 다르게 해석이 가능하다.
(오답: 개발 문장의 분석에 오류가 나타나면 많은 문서를 가공해도 추이파악에 어려움이 있다--x)
ㄴ절차 : 크롤링-필터링 -NLP -
긍/부정 감성어 추출(감성분석)

* 사회연결망 분석〉
매개중심성 측도 (betweenness centrality)
- dgree : 한 노드를 중심으로 얼마나 많은 edge가 연결 되었는 지 나타냄.
- eigenvalue centrality : 연결된 노드에 가중치를 두어 일정의 위세 정도 파악
- closeness : 각 노드들 간의 거리를 근거로 중심성 측정
- betweenness :전체 관계만 고려, 중계자 역할의 정도 측정

* 선형계획법(linear programming) :
- 최적화 방법으로 가장 많이 사용. 기대 수익,비용 고려 최대 수익 달성이 가능한 사업계획 선정 등 사업현장과 밀접한 연관

* 난수생성법(〈통계적 모의실험)
- 역변환법(inverse transform method)
- 합성법(composition method)
- 채택-기각법 (acceptance-rejection method)
(오답: 이분법--〉방정식..)

* 데이터 처리 과정에서 시간/자원의 제약 내 고품질을 내기 위한 방법 중 가장 적절한 것
- 빠르게 원시모형을 만들어 모델에 적용(prototype) --- 모델링 동일

* 상관관계가 높은 변수들이 회귀분석에 포함될 경우, 해결방안
- 중요하지 않은 변수일 경우, 해당 변수 제거
- 해당 변수의 상관관계에 따라 변수 통합
- 자료 부족이 원인일 경우, 자료 보완
(오답: 상관관계가 낮아지도로 변수값 조정--x)

* ROC(Receiver Operating Characteric) 그래프
분류모형 평가에 사용되는 그래프로
x 축은 (1-특이도), y축은 민감도

* 연관성분석=장바구니분석=서열분석
정형 데이터 마이닝 기법의 대표적인 것으로 흔히 기업의 데이터베이스에서의 상품의 구매, 서비스 등 일련의 개체 또는 사건들 간의 규칙을 발견하기 위해 사용 된다.

*군집분석 〉 계층적 방법 〉 응집분석(agglomerative) , 분할분석(divisive analysis)
ㄴ응집분석 : 군집을 나누는 방법 중 n개의 관측값을 각각 하나의 군집으로 간주. 관측값의 특성이 가까운 군집끼리 순차적으로 합해가는 방법.

! 군집분석 : 모집단에 대한 사전정보가 없는 경우, 주어진 관측값 사이의 유사성을 이용하여 전체를 몇 개의 집단으로 그룹화/ 각 집단의 성격을 파악하는 분석법.

* 연관성 분석 측도
- 지지도 (support) :
전체 거래 항목 중 항목 A와 항목 B가 동시에 포함되는 비율. 전체 거래 중 항목 A와 항목 B를 동시에 포함하는 거래가 어느정도인지 나타내주며 이를 통해 전체 구매 경향을 파악항 수 있다. 그러나 지지도는 연관규칙 A-〉B , B-〉 A가 같은 지지도를 가져 두 규칙의 차이를 알 수가 없다. 이에 대한 평가 측도로 신뢰도(confidence)가 필요하다. 신뢰도는 원인이 발생했을 때 결과가 발생할 가능성을 나타낸다.
즉, 연관규칙 A-〉B의 신뢰도는 A가 발생했을 때 B가 발생할 조건부확률이라고 생각할 수 있다.

*카이제곱(chi-square) 검정
분류조합에 따라 측정값에 유효한 차이가 발생하는 지를 검정. 명목척도로 측정된 두 속성이 서로 관련되어 있는지 측정할 때 사용.

*검정통계량 분포
F-분포: 두 집단의 평균이 같은지를 검정하려는 목적으로 "분산"이 같은지를 검정할 때
T-분포,Z-분포 : 평균 검정

*가설검정
-가설은 항상 귀무가설, 대립가설이 있다.
-검정통계량 값을 구한 후 이 값이 나타날 가능성의 크기에 의해 귀무가설 채택여부를 결정하고 이는 유의수준(siginficancr level ,a) 기준 판단
-귀무가설이 옳은데도 귀무가설을 기각하게 되는 오류를 제 1종 오류라고 한다.
-가설검정에서는 제 1종 오류의 크기를 고정시킨 후 기각역을 설정한다.

*피어슨 상관계수 확인 방법
: 단순회귀 분석에서는 r-square의 루트값에 기울기의 부호를 붙임.

*잔차 : 오차항에 대한 추정치
잔차들의 독립성은 잔차도를 통해 확인할 수없다. **
잔차의 독립성은 잔차대 순서 그래프에서 패턴을 보고 판단할 수 있다.

*상관 계수에 대한 패턴을 확인하기 위한 분석 예 : 산점도 scatter plot

*다차원척도법(MDS)
-크루스칼(kruskal)의 스트레스값을 이용하여 결과의 신뢰성, 타당성,적합성 검증
(stress 값은 응답자의 인식과 지각도 맵상 자극점들 간의 불일치 정도를 나타내는 것으로 일종의 오차의 크기를 나타내는 지수이다.)
-차원이 많아지면 stress는 개선된다.
-차원의 해석은 주관적 통찰에 주로 의존한다.
-소비자의 인식을 그림으로 표현한다.

*시계열 자료
- 정상성(stationarity)
: 모든 시점에 대한 일정한 평균, 분산을 가지며 시점 t와 s의 공분산은 시차(t-s)에만 의존하고 실제 어느 시점인지에는 의존하지 않는다.
현 시점의 자료가 과거의 자료에 의존하는 형태를 모형화 한다.
(오답: 모든 시점간에 자료는 독립이다--x)

!체크 필요
가설 오류/ 신뢰도,지지도,향상도../ 주요 함수

#자기상관함수(ACF) #부분자기상관함수(PACF) 그래프 읽기
#주성분분석 : 읽는 정보량 = 1-comulative propottion
-주성분 수를 선택할 때 고려사항
(1) 전체 변이의 공헌도(percantage of total variance)
(2) 평균 고유값(average eigenvalues)
(3) 스크리 그래프 (scree plot)

(4) 차원의 크기
(오답: 개별 고유 값의 분해 가능여부)

*데이터분석 사전 단계
- 이상치 : 이상치 무조건 제거x, 실제 오류 여부에 대해서는 어떤 통계적 이론도 설명하지 못함.
-모델의 성능은 보통 독립변수가 추가 될 수록 향상. 그러나 현 데이터 성능만을 고려하여 변수 추가 시 미래 예측하는데 부정적으로 사용될 수 있다.
-결측치(missing data) 의 숫자가 매우 적다면 이들을 제거하고 분석하는 것이 효율적이다.
-신뢰성 있는 결과를 얻기 위해 데이터의 표준화가 필요한 경우가 있다.

#이상치 판정 : 통상 평균으로부터 표준편차의 3배가 넘는 범위의 데이터.
-군집분석 : 다른 데이터들과의 거리상 떨어진 데이터로 판정
-회귀분석 : 설명변수의 동일수준의 다른 관측치에 비해 종속변수의 값이 상이한 점을 이상치로 판정
-데이터 크기순으로 나열 시 가장 크거나 가장 작은 수치
-관련 알고리즘 ESD(Extream Stedentized Deviation) ,MADM

#검정용데이터 : 구축된 모델의 과잉,과소맞춤 등에 대한 미세조정을 위해 사용.

#알고리즘 분류(지도 vs 비지도)
-비지도 : Apriori(연관규칙) , K-Means(클러스터 묶는,자율학습 일종) , SOM (신경망, 자기조직화)
-지도: C5.0 (의사결정)

#분류분석 :
- 분류작업의 목적은 새로 나타난 대상의 특징을 살펴보고, 사전 정의 된 분류의 집합들에 할당하는 모형을 만들어내는 것.
-군집분석과 달리 각 계급이 어떻게 정의되는지 미리 알아야 한다.
-분류를 의해 사용되는 데이터 마이닝 기법은 K-NN,의사결정나무모형, 신경망모형
-의사결정나무모형에서는 분할 후 생성된 노드들의 불순도 함수값의 감소가 가장 크게 일어나도록 분할이 진행된다.
-회귀나무모형 : 의사결정나무 모형에서 '목표변수가 연속형'일 때 사용하는 알고리즘

#예측분석 :
- 분류,추정과 동일하나 미래의 행위를 분류/미래의 값을 추정.
-고객행동 예측, 시계열 분석을 통해 미래의 매출 등 예측, 분류나 추정과 분리하는 이유는 예측적 모형화에 있어서는 설명변수들과 종속변수의 예측치 간의 순차적인 관계에 대한 고려가 필요하기 때문.
(오답: 고객군 분류 - 예측X 설명변수를 이용하여 세부군으로 분류Classification 하는 문제에 해당)

#오분류표
* 오분류표를 사용하여 분류분석 모형을 평가하는데 사용하는 대표적인 지표*

#정확도(Precision): T로 예측한 관측치 중 실제값이 T인 정도
ㄴ TP/TP+FP
#재현율(Recall) : 실제값이 T인 관측치 중 예측치가 적중한 정도(=민감도), 모형의 완전성을 평가하는 지표.
ㄴ TP/TP+FN = TP/P

#계절요인(〈분해시계열 분석)
: 요일마다 반복되거나 일 년 중 각 월에 의한 변화, 사분기 자료에서 각 분기에 의한 변화 등 고정된 주기에 따라 자료가 변화하는 요인.

#자기회귀이동평균모형(ARMA) (〈 시계열 모형
- 과거시점의 관측자료와 과거시점의 백색잡음의 선형결합으로 현 시점의 자료를 표현
-자기회귀AR 모형 : 과거 시점의 관측자료의 선형결함으로 표현
-이동평균MA 모형: 과거 시점의 백색잡음의 선형결합으로 표현
위 두 모형을 합치면-〉ARMA

#SVM (Support Vector Machine)은 자료들을 분리하는 초평면(hyperplane) 중에서, 자료들과 가장 거리가 먼 초평면을 찾는 방법. 여백(margin)을 최대화하여 일반화 능력을 극대화하는 방법이라 할 수 있다.
(일반 신경망 알고리즘은 오류율을 최소화하는데 목적이 있다.)

-특징
(1) 주어진 문제에 대해 자동으로 최적의 커널을 선택하는 알고리즘 없다.
(2) 커널과 과년된 매개변수를 다양하게 설정 및 성능 실험/그 중 가장 뛰어난 값을 선택하는 휴리스틱한 방법 사용
(3)최적파라미터 설정 방법 없다.
(4)직선으로부터 가장 가까운 샘플까지의 거리가 동일하다.

#군집분석 거리 정의 (〈계층적 군집) :
- 두 개체 간의 거리(or 비유사성)에 기반해 군집을 형성에 나가므로 거리에 대한 정의 필요.
- 유클리드 거리(Euclodian)
- 맨하튼 (Manhattan) : 변수들이 연송형, 이상치를 제거할 수 없는 경우 사용할 수 있는 로버스트(Robust)한 측도, 시가(city-block) 거리
-민코우스키(Minkowski)거리
----------(위) 수학적 거리 ---(아래) 통계적 거리 : 변수간 상관성 동시 고려----------
-표준화(standardized) 거리 : 관측단위의 영향을 없애기 위해 일반적으로 사용
-마할라노비스(Mahalanobis distance) 거리 :변수들 간 상관관계가 있다고 판단될 때 사용

*데이터 분석 요건 도출 간계
- 이슈리스트 작성
- 핵심 이휴 정의
- 이슈 그루핑
- 해결 방안 정의

#모델성능 평가 기준
- 정확도(정분류율, accuracy) = TP+TN/P+N
- 정밀도(precision) = N/TN
- 민감도(sensitivity) =TP/P
-error rate = FP+FN/P+N

#연관규칙 : 판매시점에서 기록된 거래와 품목에 대한 정보 필요. 특정 고객들의 인구통계학적 자료를 비롯 기타정보 불필요

*연관규칙 사용 측도
-지지도(support) : 전체 거래 중 품목 A,B가 동시에 포함되는 거래의 비율 (전체 거래 중 A,B가 동시에 포함되는 거래의 정도)
전체 구매 경향 파악에 용이, 그많큼 많은 양, 불필요한 분석을 대폭 줄일 수 있다.

-신뢰도(confidence) : 품목 A가 포함된 거래 중, 품목 A,B를 동시에 포함하는 거래일 확률 (연관성 정도)
A를 구매한 사람이 B도 구매할 경우, A->B (B->A와는 다름!)

-향상도(lift) : 품목B를 구매한 고객 대비 품목 A를 구매한 후 품목 B를 구매한 고객에 대한 확률, 연관 규칙 A->B는 품목 A,B의 구매가 서로 관련의 없을 경우 P(B|A)=P(B) ,향상도는 1. 향상도가 1보다 크면 예측에 있어 우수하다고 평가 가능. (1보다 작을 경우 음의 상관관계)
양의 상관관계 일 경우, 'B를 구매할 확률보다 A 구매 후 B를 구매할 확률이 높다'고 해석할 수 있다.
$지지도=₩ P₩left(A₩cap B₩right)₩ ₩ =₩frac{A와₩ ₩ B가₩ 동시에₩ 포함된₩ 거래수}{전체₩ 거래수}$지지도= P(A∩B) =
A와 B가 동시에 포함된 거래수
전체 거래수

$$
$ 신 뢰 도 = ₩ f r a c { P ₩ l e f t (A ₩ c a p B₩right)}{P₩left(A₩right)}=₩ ₩frac{A와₩ B가₩ 동시에₩ 포함된₩ 거래수}{A를₩ 포함하는₩ 거래₩ 수}$신뢰도= P(A∩B)
P(A)
=
A와 B가 동시에 포함된 거래수
A를 포함하는 거래 수

$$
$ 향 상 도 = ₩ f r a c { P ₩ l e f t (A ₩ c a p B₩right)}{P₩left(A₩right)₩cdot P₩left(B₩right)₩}=₩ ₩frac{A와₩ B를₩ 포함하는₩ 거래수}{A를₩ 포함하는₩ 거래수₩ ₩cdot ₩ B를₩ 포함하는₩ 거래수}$향상도= P(A∩B)
P(A)·P(B)
=
A와 B를 포함하는 거래수
A를 포함하는 거래수 · B를 포함하는 거래수

*데이터 처리 : 데이터 분석을 위해 분석 방법에 맞게 데이터를 수집/변형하는 과정
-정체는 표준화와 잘못된 데이터를 수정하는 작업이 필요하다.
-데이터 처리과정은 많은 시간과 노력이 필요하다. 제일 좋은 방법은 빠르게 원시모형(portotype)을 만드는 것이다.
-데이터마트는 조회를 이용한 분석인 OLAP,리포팅 등에도 활용할 수 있어 분석업무의 운영적 측면에서의 활용성이 높다.
-원하는 데이터 형태로 가공하는 과정은 분석 결과의 품질과 성능에 큰 영향을 미친다.

#계통추출법
단순랜덤추출법이 변형된 형태로 N개의 모집단의 원소들을 n개의 계통으로 나눈 후 각 계통에서 표본을 랜덤하게 추출하는 방법 .

#구간척도
측정 대상이 가지고 있는 속성의 양을 측정. 측정결과가 "숫자"로 표현되나 해당 속성이 전혀 없는 상태인 절대적 "0"이 없다. (온도, 지수)

*그룹별 비교 중 "표본에 의한 비교"시 자료 구조
-각 그룹에서의 관측값들은 각 모집단에서의 랜덤표본이다.
-각 처리를 적용할 실험단위를 랜덤하게 하는 과정은 랜덤화 과정으로 연구에서 가장 기본,핵심적인 작업이다.
-서로 다른 그룹에서의 관측값들은 독립적으로 관측된 것이다.

*정상성(<시계열 자료)
- 시계열 자료가 평균이 일정하고, 분산이 시점에 의존하지 않으며, 공분산은 시차에만 의존하고 실제 어느 시점에는 의존하지 않는 다.
는 세 가지 조건을 만족할 경우를 말한다.

*데이터 마이닝 추진 단계
(1)목적 설정: 데이터마이닝의 명확한 목적을 설정
(2)데이터 준비 : 필요에 따라 웹로그 데이터나 sns 데이터 활용
(3)가공 : 모델링 목적에 따라 목적 변수를 정의하고 , 필요한 데이터를 데이터 마이닝 소프트웨어에 적용할 수 있도록 함.
(4)기법 적용 : 앞 단계를 거쳐 준비한 데이터와 데이터 마이닝 소프트웨어를 활용해 목적하는 정보를 추출
(5)검증 : 마이닝으로 추출한 정보를 검증

*사회연결망분석(SNA : Social Network Analysis)
개인와 집단 간의 관계를 노트/링크로 모델링. 그것의 위상구조와 확산 및 진화과정을 계량적으로 분석하는 방법론으로
집합론적 방법, 그래프 이론을 이용한 방법, 행렬을 이용한 방법 등이 있다.

#이산형 확률변수 : 사건의 확률이 그 사건들이 속한 점들의 확률의 합으로 표현할 수 있는 확률변수
ㄴ 종류 : 베르누이 ,이항분포(binomial distribution) ,기하분포(geomerric-), 다항분포(multinomial-), 포하송분포(poisson-)
#연속형 확률변수 : 균일분포(uniform-), 정규분포(normal-), 지수분포(exponential-)

#신뢰구간
-신뢰수준이 올라가면 신뢰구간은 길어진다.
-관측치의 수가 늘어나면 신뢰구간의 길이는 줄어든다.
-점추정의 정확성을 보완한다.
-주어진 신뢰수준 하에서 무수가 특정한 구간에 있을것이라 선언하는 것이다.

#상관계수
-등간척도 이상으로 측정되는 두 변수 간의 상관관계 측정 : 피어슨 상관계수
-서열척도인 두 변수간의 상관관계를 측정: 스피어만 상관계수

#회귀분석 가정
- 독립변수와 종속변수는 선형관계
-독립변수의 모든 값에 대하 오차들의 분산 일정
-관측치의 잔차끼리 상관이 없음
-잔차가 정규분포를 따름(오답: 종속변수)

#시계열 데이터
-추세(trecd) : 한 시점에 서 다음 시점으로의 전반적인 패턴 변화
-계절변동(seasonality) : 짧은 기간 동안의 주기적인 패턴
-수준(level) : 시계열의 평균값
-잡음(noise) : 무작위적인 변동, 보통 알 수 없는 이유로부터 발생

#Q-Q plot (qqplot) : 45도 직선과 가까울 수록 정규분포에 따름.
(출처: 위키피디아)

#결측값
-R에서는 "NA"
-수학적으로 불가한 수 표시 "NaN"
-데이터 유형과 자료의 길이도 0, 비어있는 값 "NULL"

*데이터마이닝 수행 작업 관련
-예측은 분류 또는 추정과 동일하지만, 미래의 행위를 분류/값을 추정한다는 점에서 차이가 있다.
-연관성규칙 작업은 어떤 일이 함께 발생할지를 판단하는 것이다.
-군집화는 이질적인 사람들이 모집단으로부터 다수의 동질적인 하위집단 혹은 군집들로 세분화 하는 작업이다.
-분류는 이산형, 추정은 연속형 값을 가지는 결과를 다룬다.

#잔차분석(analysis of residual)
회귀분석 시 직선을 관계를 어느정도 설명할 것이라 판단되면, 잔차의 선형성, 등분산성,독립성, 정규성 등을 검토하게 된다.
잔차를 이용하여 가정을 검토하는 것을 말한다.

#CV(coross validation) : 데이터양이 충분하지 않은 경우 모형을 평가하는 방법

의사결정나무 분석 활용 예
- 고객을 신용도에 따라 우량 또는 불량으로 구분
- 다수의 예측변수 중 목표변수에 큰 영향 미치는 변수를 탐색
- 웹사이트 회원들이 가장 잘 반응하는 이메일마케팅모델 구축
(오답: 고객 속성에 따라 고객을 여러 개의 배타적인 집단으로 구분---〉 군집분석"집단" 키워드)

군집분석 :
- 정형데이터마이닝 기법. 타깃변수 없는 데이터에서 우리가 몰랐던 숨은 유용한 데이터 구조 찾는 자율학습기법에 속함.
- 관측치 특성에 따라 여러 배타적 집단으로 나누는 방법.
- linkage clustering : 계층적 군집분석 방법의 하나이다.
- K-means clustering : 알고리즘 수행 과정 중 한 개체가 속해있던 군집에서 다른 군집으로 이동/ 재배치가 가능하다.
ㄴ 초기값 선택이 최종군집 선택에 영향을 미친다. 초기 군집수 결정이 어렵다, 각 데이터를 거리가 가장 가까운 seed가 있는 군집으로 분류한다.
-군집화 dendogram 은 "계층적 군집분석"에서 가능

변수선택(〈회귀분석)
전진선택법 : 중요하다고 생각하는 설명변수를 차례로 모형에 추가
후진제거법: 모든 설명변수를 포함한 모형에서 출발. 종속변수의 설명에 가장 적은 영향을 주는 변수부터 제거
단계별방법: 전진선택법에 의해 번수 추가, 새로 추가된 변수에 기인해 기존변수의 중요도가 약화되면 제거
모든 가능한 조합의 회귀분석: 모든 가능한 독립변수들의 조합에 대한 회귀모형을 고려, AIC 또는 BIC 의 기준으로 가장 적합한 회귀모형 선택

*분석기법 알고리즘

-앙상블 : 여러 개의 모형을 결합해 개별 모형보다 좋은 예측 성능을 얻는 분석 기법
(알고리즘 ; bagging, boosting, randome forest)
-차원축소: 고차원의 자료를 변수들 간의 선형, 비선형 결합으로 생성된 기존 변수들보다 적은 수의 새로운 변수들로 근사시키는 방법
-고차원 회귀분석: 독립변수의 개수가 관측치 개수에 비례해서 증가하거나 매우 많은 경우
종속변수에 영향을 미치는 적은 개수의 독립변수의 선형 결합으로 종속변수를 예측하는 기법
-최적화 : 주어진 제약조건 하에서 달성하고자 하는 목표를 이루기 위한 의사결정 문제를 모형화하고 이에 대한 해를 구하기 위한 기법

*벡터의 인덱스는 양의 정수, 음의 정수 함께 쓸 수 없다.(예. fruit[-2:3]--X)

#모형개발을 위한 준비작업 순서
: 데이터 추출 - 데이터 정제- 데이터 파생- 데이터 분할

카이제곱검정:
- 동질성 검정이나 독립성검정을 하기 위한 검정법.
(귀무가설 예. "흡연빈도와 운동량은 서로 독립이다.")

다중공선성(Multicollinearity) :
- 독립변수들 간에 높은 선형관계가 존재 하는 경우
- 중요하지 않은 변수는 제거
- 분산팽창요인(variance inflation factor) 로 값이 10이 넘을 경우 다중공선성 문제가 있음으로 판단
- 결정계수 값은 높으나 독립변수의 P-값이 커서 개별 인자들이 유의하지 않을 경우 다중공선성 의심
(오답: 상관관계가 낮아지도록 변수값 조정)

어프라이어리(apriori) 알고리즘
: 연관분석을 수행하기 위해 빈발 아이템 집합과 연관규칙이라고 하는 두 가지 형태로 표현하는, 연관성 분석을 수행하는 대표적인 1세대 알고리즘
최소지지도를 갖는 연관규측을 대표적인 방법.
최소지지도보다 큰 집합만을 대상으로 높은 지지도를 갖는 품목집합을 찾는 것.

시뮬레이션
실제상황을 수학적으로 모델화 하고 그 모델을 컴퓨터에 프로그램으로 저장한 후, 일어날 수 있는 가능한 모든 상황을 입력함으로써
각각의 경우에 어떤 결과가 도출되는지 예측하는 것.

*의사결정나무모형 구축 시 최적의 분할 변수를 선택할 때 , 불순도 척도(impurity measure) 를 사용한다.
불순도 척도의 종류는 엔트로피, 지니계수, 분류오류율이 있다.

*다중회귀분석 중 종속변수를 설명하는데 더 중요한 독립변수〉
- 표준화 자료로 추정한 계수(coefficient)

#NaN : 0/0 , Inf-Inf (수학적으로 불가한 수를 표시할 때 사용)
#회귀나무, 회귀나무모형 (regression tree) : 연속형 타깃변수(목표변수)를 예측하는 의사결정 나무
#corpus : 텍스트 마이닝에서 더 이상의 추가적인 절차 없이 데이터마이닝 알고리즘 실험에서 활용될 수 있는 상태, 문서들의 집합
데이터마이닝 절차 중 데이터의 정체, 통합,선택,변환의 과정을 거친 구조화된 단계

#t-통계량 : 주어진 신뢰도 하에서 모수가 특정한 구간에 있을 것으로 선언 , 모분산이 알려져 있지 않은 경우 사용

#분류분석 : 고객이 여러 속성(나이,성별,직업,과거 구매행태 등)을 이용하여 해당 고객의 이탈 여부를 예측
관측치가 미리 정의된 어떤 그룹에 속하는지 예측하는 데 사용

#상관분석
- 두 변수의 상관관계를 알아보기 위해 상관계수를 이용한다.
-상관계수의 값은 항상 -1과 +1사이에 있으며, +1에 가까울수록 양의 상관관계를 나타낸다.
-상관계수의 값이 0에 가까운 것은 두 변수가 약한 상관성을 가진다고 보고, 아무 관계가 없는 것은 아니다.
-상관 분석은 연관 정도만 나타낼 뿐 인과관계는 나타내지 않는다.

* 다중회귀분석 실시 이후, 모형의 적절성을 확인하기 위한 체크 사항
- F-통계량 확인을 통해 모형이 통계적으로 유의미한지 확인
- T-통계량, P-값 등을 통해 유의미 한지 확인
- 잔차그래프를 그려 모형이 데이터를 잘 적합하고 있는지 확인
- 상관계수로 변수의 상관관계와 그 정도 파악
- 결정계수는 0~1의 값을 가지며, 높은 값을 가질수록 추정된 회귀식의 설명도가 높다.(오답: 설명력은 상관계수가 아니라 결정계수!)

*변수 축소- 주성분분석의 주성분 개수를 결정하는 알고리즘에 대한 설명
-기준보다 큰 고유치의 개수를 이용
-표본 공분산 행렬의 고유치 이용
-전체 분산을 설명하는 비율이 기준치를 넘는 주성분의 수를 이용
-변수들의 선형결합으로 이루어진 주성분은 서로 독립
-주성분분석은 다변량 자료 분석에 이용하는 방법으로 독립변수에 사용한다.(종속변수x)
-차원 축소를 통해 자료의 시각화에 도움을 줄 수 있으면 차원이 축소된 주성분으로 회귀분석에도 적용 가능하다.
-변수들의 선형결합으로 이뤄진 주성분은 서로 독립, 기존 자료보다 적은 수의 주성분들로 기존 자료의 변동을 설명한다.
-부분최소제곱법이란 독립변수와 종속변수의 변동성을 가장 잘 설명하는 새 변수를 설정하고, 이들의 관계를 통해 종속변수와 독립변수의 인과관계를 분석하는 방법이다.

*통계적 추론 〉 모집단 모수에 대한 검정 방법
1. 모수적방법(parametric method)
-관측된 자룔로 구한 표본평균과 표본분산 등을 이용해 검정을 실시한다.

2. 비모수적방법(nonparametric method) :
부호검정, 순위합검정, 만-위트니U검정
- 자료가 추출된 모집단의 분포에 대해 아무 제약을 가하지 않고 검정을 실시하는 방법
- 관측된 자료가 특정분포를 따른다고 가정할 수 없는 경우에 이용
- 관측된 자료의 수가 많지 않거나 자료 자체가 개체간의 서열 관계를 나타내는 경우에는
관측된 자료가 주어진 분포를 따른다는 가정을 받아들일 수 없는 경우에 이용하는 검정법이다.

#선형계획법 (Linear Programming) : 생산량, 비용, 인원등의 데이터가 1차 함수로 주어졌을 때 목적함수에 대해 최적의 해를 얻는 방법.
자원을 용도에 맞체 효율적으로 배분하는 기본적 문제를 해결하는 데 사용되는 최적화 기법으로 기업에서 많이 사용함.

#표본 : 조사하고자 하는 대상 집단 전체인 모집단 모두를 조사하는 것은 비용,시간 많이 소요.
모집단을 적절히 대표할 수 있는 일부 원소들을 뽑아 관찰 파악하여 모집단을 유추, 이 때 추출한 모집단의 부분집합

*시계열 구성요소
추세요인 : 자료가 어떤 특정한 형태를 취할 때 추세요인이 있다고 한다.
계절요인: 고정된 주기에 따라 자료가 변화할 경우
순환요인: 경제적이나 자연적인 이유 등 잘 알려진 주기를 가지고 자료가 변화할 때
불규칙요인 : 추세,계절,순환요인으로 설명할 수 없는 회귀분석에서 오차에 해당하는 요인

#신경망(NN) 특징
-변수의 수가 많고 입력변수와 출력변수가 복잡한 비선형형태를 가질 때에도 다른 분류모형보다 비교적 정확도가 우수
-훈련용 데이터에서는 만족스러운 결과이나, 실제 적용에서는 분류가 정확하지 않은 모형의 과대적합(overfitting)현상을 일으키는 경우 종종있다.
-훈련용 데이터에 잡음이 있더라도 민감한 반응을 보이지 않는다.
-분류결과에 대해 왜 그렇게 되었는지 설명할 수 없는 블랙박스 형태로 해석할 수 없다는 단점. (해석용이--〉 의사결정나무 모형)

VS 의사결정나무 모형
-모형의 결과를 누구에게나 설명이 용이해 해석력이 우수
-무형구축하는 방법이 계산적으로 복잡하지 않는다.
-비정상적인 잡음 데이터에 민감하지 않게 분류 가능
-불필요한 변수가 많아지면 나무모형 크기가 커질 수 있어 불필요한 변수를 제거하는 것이 좋음.

#사회연결망 분석 〉 네트워크 중심성 측정방법
(1)연결정도 중심성(degree centrality) : 한 점에서 직접적으로 연결된 점들의 합으로 얻어지며, 한 점의 포인트 중심성을 측정
(2)근접 중심성(closeness-) : 각 노드간의 거리를 근거로 직접적으로 연결된 노드 및 간접 연결된 모든 노드의 거리를 합산해 중심성 측정
(3)매개 중심성(between-) : 네트워크 내에서 한 점이 담당하는 매개자 혹은 중재자 역할의 정도로서 중심성을 측정
(4)위세 중심성(eigenvector-) : 연결된 노드의 중요성에 가중치를 둬 노드의 중심성을 측정

*연관규칙 측정 예제
빵-〉 우유
- 지지도 : 빵+우유 동시 포함/ 전체 거래수
- 신뢰도 : 빵+우유 동시 포함/ 빵 포함 거래수
- 향상도 : 빵+우유 동시 포함/ 빵 포함거래수 * 우유 포함거래수

#스태밍(stemming) : 텍스트 마이닝의 전처리 과정 중 변형된 단어형태에서 접사(affix) 등을 제거하고 그 단어의 원형 또는 어간을 찾아내는 것
#term-documnet matrix(단어 문서행렬) : 텍스트마이닝 전처리과정 이후 문서번호와 단어 간의 사용여부 또는 빈도수를 이용해 만들어진 행렬

*사회연결망분석 〉 네트워크 구조 파악하는 기법 : 밀도, 집중도, 구조적 틈새

#중앙값 : 데이터를 크기순으로 나열할 때 가장 중앙에 위치하는 데이터 값(오답: 관측순으로 나열할 때)

- 계층적 군집방법 문제 풀이 시, 가장 가까운 거리부터 가장 먼저 군집 형성

*R함수 패키지 #reshape : 단지 melt, cast만을 사용하여 데이터를 재구성하거나 밀집화된 데이터를 유연하게 생성해 준다. 데이터 정보들을 그대로 유지한 상태로 재정렬 수행.
- melt()
melt(df, id=c("Month", "Day"), na.rm=T)
Month, Day 기준으로 뽑아줌
#R패키지- 분류
klaR(분류&시각화) , party(의사결정나무분석)

#buzz분석(=buzz량 분석)
: 비정형 데이터마이닝 분석 방법 중 특정기간별 발생 문서량(예. 온라인에서 언급된 횟수)의 추이를 분석하는 것

*데이터프레임 명령어
(1)결측치 포함, 관측치 제거 - na.omit(data.frame이름)
(2) NA 포함한 인덱스를 논리벡터, 논리매트릭스로 추출 - is.na()

(3) 함수 내에서 사용되는 옵션, NA무시 -na.rm=
(4)결측치 제거 - !is.na()

*R명령어 : 그룹별 요약(sum,mean..)
data.table, sqldf, aggregate
(오답: melt : 불가)
*R 박스플랏 스크립트
boxplot(변수x1~변수xi, 데이터명)
*데이터 유형
data.frame : 숫자형+문자형 함께 포함 할 수 있다. (오답: 없다)
matrix : 차원을 가진 벡터. 숫자형 원소와 문자형 원소를 함께 포함할 수 없다.
list : 각 요소는 서로 다른 모드의 객체를 포함할 수 있다. & 각 요소는 [[]]로 접근 가능하다.

* paste (R 함수) : 문자열 혹은 벡터들을 지정된 구분자를 사용해 결합한다.

*pnorm :
"p"가 누적분포함수를 나타냄.
ㄴ d: 확률값 , p:누적확률값 , q:백분위수 , r:난수발생
*R코드 출력 예제 및 결과
x<-matrix(c(1:12),3,4)
; [,1] [,2] [,3] [,4]
[1,] 1 4 7 10
[2,] 2 5 8 11
[3,] 3 6 9 12

m<- matrix(1:6,nrow=3)
; [,1] [,2]
[1,] 1 4
[2,] 2 5
[3,] 3 6
- cast(): 원하는 형태로 계산 또는 변형
cast(df, Day~Month~variable, func)
cast(df, Day~Month+variable,func)
#pairs() : 산점도 행렬 도식화 할 때 사용하는 함수
#ggmap() : 지도 기반 시각화
#hist(): 히스토그램
#boxplot(): 상자그림

*R의 데이터 오브젝트에 대한 설명
-벡터에서 모든 원소는 같은 모드를 갖고 있어야 한다.
-리스트에서 원소들은 다른 모드여도 된다.
-행렬은 차원을 가진 벡터
-데이터프레임은 리스트로 구현된 구조이다. (테이블로 된 구조 -오답)
*데이터프레임 추출 방법
-변수를 벡터로 추출 : [[index]].[["변수명]],$변수명 ..
-변수를 데이터프레임으로 추출 : [index], ["변수명"]

*통계량의 "열" 기준 평균 산출
apply(data,2, mean) --> 행 구할 대는 apply(data,1,mean)
sapply(data,mean)
lapply(data,mean)
colMeans(data)

*주성분분석 : 자료의 Cumulative Proportion 확인

*ARIMA모형
AR1 계수 -ar1 아래(알파를 구하시오) , MA 계수-ma1 아래

#set.seed : R함수, 난수가 항상 동일하게 발생되도록 초기화하는 함수.

#sna패키지 : 소셜네트워크분석
ㄴgden() 밀도 계산함수(=총 연결선수/가능한 총 수)
#의사결정나무모형 R패키지
: rpart, party, maptree

#df : 자유도 (총 자료의 수-2)
! 오답- 전체 자료의 개수