문항 1. 데이터 분석과 빅데이터 분석의 차이점에 4가지에 대해서 설명하시오

데이터 분석

* (1) 데이터 분석은 원시 데이터를 실행 가능한 인사이트로 변환함
* (2) (1)을 토대로 의사 결정권자에게 실행 가능한 결론에 도출하게 끔함.
* (3) 데이터에서 얻는 인사이트는 차트, 그래프 시각적인 형태로 표현이 가능함
* (4) 원시 데이터에서 얻은 정보는 회사, 조직 수익 증대를 도모할 수 있음.

빅데이터 분석

* (1) 규모가 매우 큰 데이터를 대상 함
* (2) 로그데이터(log data)나 구매기록 등 정형 데이터 뿐 아니라 소셜 미디어, 위치, 센서 등 반정형·비정형데이터까지 분석대상에 포함함
* (3) 다양한 데이터의 관계를 동시에 가능한 빨리 처리할 수 있는 새로운 컴퓨팅 기술을 적용함
* (4) 다양하고 신뢰할 수 있는 분석결과를 제시하여 가치를 창출하는 데이터 처리방식임

상기 명의 내용으로 봤을 때, 데이터 분석과 빅데이터 분석은 규모의 차이가 가장 크다고 볼 수 있음.

문항2. 데이터 전처리 기술 3가지가 무엇인지 명칭을 작성하고, 그 개념에 대해 설명하시오.(필요시, 사용되는 세부적인 기술을 설명하시오.

데이터 전처리

* 1) 데이터 정제 -> 2) 결과 측정값 처리 -> 3) 이상값 처리 -> 4) 분석 변수의 처리
* 1. 데이터 정제

오류의 원인분석 -> 정제의 대상 선정 -> 정제의 방법을 결정으로 볼 수 있음

* 2. 결과 측정값 처리

완전 무작위 결측 -> 무작위 결측 -> 비무작위 결측

* 3. 이상값 처리

개별 데이터 관찰 -> 통계값 -> 시각화 -> 머신 러닝 -> 마할라노비스 거리 -> LOF -> iForest

* 4. 분석 변수의 처리

전진 선택법 -> 후진 선택법 -> 단계적 선택법

**디테일 설명은 하기에 기입함**

**데이터 정재**

1. 데이터 전처리의 중요성
   1. 데이터 분석과정에서 데이터 전처리는 반드시 거쳐야 하는 과정
   2. 결과에 직접적인 영향을 끼치므로 전처리는 반복적으로 수행해야한다.
   3. 데이터 전처리는 데이터 정제 -> 결측값 처리 -> 이상값 처리 -> 분석 변수 처리
2. 데이터 정제의 개념 : 결측값을 채우거나 이상값을 제거하는 과정을 통해 데이터의 신뢰도를 높이는 작업.
3. 데이터 정제 절차 : 오류 원인 분석 -> 정제 대상 선정 -> 정제 방법 결정
   1. 오류 원인 원인
      1. 결측값 : 필수적인 데이터가 입력되지 않음
      2. 노이즈 : 실제는 입력되지 않았지만 입력된것으로 판단된 것.
      3. 이상값 : 데이터 범위에서 튀는값.(기준에서 많이 벗어난 값)
   2. 정제 대상 선정
      1. 모든 대상을 기준으로 선정하는 것이 기본
      2. 특히, 데이터의 품질을 떨어트리는 데이터를 더 많이 정제해야한다.
      3. 내부데이터 보단 외부데이터,  정형데이터 보단 비정형,반정형 데이터가 품질 저하 위협에 많이 노출
   3. 정제 방법 결정
      1. 삭제 : 무작위적인 삭제는 오류를 일으킬 수 있으므로 가급적으로 피한다.
      2. 대체 : 오류 데이터를 최빈값,중앙값,평균값 등으로 대체.
      3. 예측값삽입 : 회귀식등을 이용하여 예측값을 삽입
4. 데이터 정제 기술
   1. 일관성 유지를 위한 정제기술
      1. 변환 : 다양한 형태로 표현된 값을 일관된 형태로 변환하는 작업
      2. 파싱 : 정제 규칙을 적용하기 위해 유의미한 최소 단위로 분할하는 작업.
      3. 보강 : 변환,파싱,수정,표준화 등을 통한 추가 정보를 반영하는 작업.
   2. 데이터 정제 기술 : 분산 처리 기술을 기반으로 데이터를 정제 성능 보장을위해 인메모리 기반 컴퓨팅 기술사용
      1. ETL ; 수집 대상 데이터를 추출,가공,변환하여 DW,DM 에 저장하는 기술.
      2. MapReduce : 데이터를 추출하는 Map, 추출한 데이터를 중복이 없게 처리하는 Reduce 기술로 구성     (구글에서 대용량 데이터  세트를 처리하거나 생성하기 위한 목적으로 만든 SW 프레임워크)
      3. Spark.Storm : inMemory 방식의 데이터 처리 방식
         1. in memory : 디스크에 최적화된 데이터 베이스 보다 더 빠른 접근과 처리가 가능하도록 메인 메모리에 설치되어 운영되는 데이터 처리 방식을 말한다.
      4. CEP : 실시간으로 발생하는 이벤트 처리에 대한 결과값을 수집하고, 처리하는 기술
      5. Pig :  대용량 데이터 집합을 분석하기 위한 플랫폼
      6. Flumn :  로그 데이터를 수집하고 처리하는 기법, 실시간에 근접하게 데이터를 전처리,수집
5. 데이터 세분화
   1. 개념 : 데이터를 기준에 따라 나누고, 선택한 매개변수를 기반으로 유사한 데이터를 그룹화 효율적으로 사용함.
   2. 방법
      1. 계층적 방법 : 사전에 군집수를 정하지 않고 단계적으로 단계별 군집결과를 산출하는 방법
         1. 응집분석법 : 각 객체를 하나의 소집단으로 간주하고 단계적으로 유사한 소집단을 합쳐 새로운 소집단을 구성하는 방법
         2. 분할분석법 : 전체 집단으로 부터 시작하여 유사성이 떨어지는 객체들을 분리하는 방법.
      2. 비계층적방법 : 군집을 위한 소집단수를 정해놓고, 각 객체 중 하나의 소집단으로 배정하는 방법
         1. 인공신경망 : 생물학적 신경망에 영감을 얻어 통계학적 학습모델
         2. k-means : K개의 소집단의 중심좌표를 정하여 각 객체와 중심간의 거리를 통해 군집을 나눔.

**데이터 결측값 처리**

1. 결측값의 종류   
   1. 완전 무작위 결측 ( MCAR ) : 발생한 결측값이 다른 값들과 전혀 연관성이 없는 값
   2. 무작위 결측 (MAR) : 누락된 자료가 특정 변수와 관련되 일어나지만, 그 변수의 결과는 관계가 없는 경우.                                   누락이 전체 정보로 설명이 가능함.
   3. 비무작위결측 (MNAR) : 누락된 값이 다른 변수와 연관이 있는 경우
2. 결측값 처리 절차
   1. 결측값 식별 : 다양한 형태로 표현되어 있는 결측값의 현황을 파악한다.
   2. 결측값 부호화 : NA(기록되지 않은값), NULL(값이 없음), NaN(수학적으로 정의되지 않은 값), inf(무한)
   3. 결측값 대체 : 결측값의 자료형에 맞게 대체 알고리즘을 통해 결측값을 처리
3. 데이터 결측값 처리 방법
   1. 단순 대치법 :  결측값을 가진 자료 분석에 사용하기 쉽고,  통계적 추론에 사용된 통계량의 효율성 및 일치성 등의 문제를 부분적으로 보완해준다.
      1. 단순대치법의 종류
         1. 완전 분석법 : 불완전 자료는 모두 무시, 완전한 자료만 사용하여 분석, 분석은 쉽지만 효율성이 상실되고, 통계적 추론의 타당성 문제가 발생한다.
         2. 평균 대치법 : 자료의 평균값으로 결측값을 대치해서 불완전한 자료를 완전한 자료로 만드는 방법
         3. 단순 확률 대치법: 평균 대치법으로 추정된 통계량을 기반으로 확률적으로 값을 선정하여 대치
      2. 단순 확률 대치법의 종류
         1. 핫덱(Hot-Deck)대체 : 무응답 -> 비슷한 성향을 가진 응답자의 응답으로 대체, 표본조사에서 사용
         2. 콜드덱(Cold-Deck)대체 :  현재 진행중인 연구가 아닌 과거 혹은 외부의 연구에서 결과값을 참조
         3. 혼합방법
   2. 다중 대치법 : m번의 대치를 통해 m개의 가상적 완전한 자료를 만들어 분석하는 방법
      1. 다중 대치 적용 방식
         1. 대치 : 결측자료의 예측분포 또는 사후분포에서 추출된 값으로 결측값을 대치, 베이지안 방법사용
         2. 분석 : 같은 예측 분포로부터 D개의 대치 표본을 구하여 원하는 분석을 각각 수행
         3. 결합 : 모수 세타의 점추정과 표준오차의 추정치를 D개 구한 후 이들을 결합하여 하나의 결과를 제시.

**데이터 이상값 처리**

1. 데이터 이상값 : 관측범위에서 많이 벗어난 아주 작은값 또는 아주 큰 값을 의미한다.                                                             입력,데이터처리 오류 등의 이유로 특정 범위에서 벗어난 데이터값을 의미한다.
2. 데이터 이상값 발생원인
   1. 데이터 입력오류 : 전체 데이터 분포와 비교하여 쉽게 구분가능
   2. 측정오류
   3. 실험오류 : 실험조건이 동일하지 않아 오류
   4. 고의적인 이상값 : 자기 보고식에 의한 오류
   5. 표본추출에러 : 데이터 샘플링 과정에서 오류가 발생.
3. 데이터 이상값 검출 방법
   1. 개별 데이터 관찰 :  무작위 추출 혹은 전체 데이터 추이나 특이사항 관찰하여 이상값 검출.
   2. 통계값 : 통계(평균,중앙,최빈)값 과 데이터 분산도(범위,분산)을 활용한 이상값 검출
   3. 시각화 : 데이터 시각화를 통한 지표 확인으로 이상값 검출
   4. 머신 러닝 기법 : 데이터 군집화를 통한 이상값 검출
   5. 마할라노비스 거리 : 데이터 분포를 고려, 관측치가 평균으로 부터 벗어난 거리를 측정
   6. LOF : 관측치 주변의 밀도와 근접한 관측치 주변 밀도의 상대적인 비교를 통해 이상값을 탐색
   7. iForest : 관측치,관측치주변 밀도에 의존하지 않고 Decision Tree를 이용하여 이상값 검출
4. 통계 기법을 이용한 이상값 검출
   1. ESD : 평균으로 부터 3표준편차 떨어진 값을 이상값으로 판단한다.
   2. 기하평균을 활용 : 기하평균으로 부터 2.5 표준편차 떨어진 값을 이상값으로 판단.
   3. 사분위 수를 활용 : 1,3 사분위를 기준으로 사분위간 범위(Q3-Q1)의 1.5배 떨어진 값을 이상값으로 판단
   4. 표준화점수(Z) : 평균 오메가와 표준편차가 알파인 정규 분포를 따르는 관측치들이 자료의 중심으로 부터 얼마나 떨어져 있는지를 기준으로 이상값을 검출
   5. Q검정 : 오름차순으로 정렬된 30개 미만의 데이터의 범위에서 관측치간의 차이를 통해 이상값 여부 검증
   6. T검정 : 정규분포를 만족하는 단변량 자료에서 이상값을 검정하는 방법
   7. 카이제곱검정 : 데이터가 정규분포를 만족하나, 자료의 수가 적을때 사용하는 방법
5. 시각화를 통한 이상값 검출
   1. 확률밀도함수
   2. 히스토그램
   3. 시계열차트
6. 머신러닝기법을 통한 이상값 검출
   1. 주어진 데이터를 K개로 묶는 알고리즘, 각 클러스터와 거리 차이의 분산을 최소화 하는 방법
7. 마할라노비스 거리
   1. 데이터의 분포를 고려한 거리, 관측치가 평균으로 부터 얼마나 벗어났는가
   2. 이상값 탐색을 위해 고려되는 모든 변수간에 선형관게를 만족하고, 각 변수들이 정규 분포를 따르는 경우 적용
8. LOF(Local Outlier Factor)
   1. LOF는 관측치 주변의 밀도비교를 통해 이상값을 탐색
   2. 각 관측치에서 k번째 근접이웃까지의 거리를 산출하여 해당 거리 안에 포함되는 관측치의 개수의 역수
9. iForest
   1. 관측치 사이의 거리 또는 밀도에 의존하지 않고, 데이터 마이닝 기법인 의사결정 나무를 이용하여 이상값 탐지.
   2. 데이터의 평균적인 관측치와 멀리 떨어진 관측치일수록 적은 횟수의 공간 분할을 통해 고립시킬 수 있다.
10. 데이터 이상값 처리 : 반드시 제거해야 하는 것이 아니므로 이상값을 처리할지는 분석의 목적에 따라 적절한 판단이 필요하다.
    1. 삭제 : 이상값으로 판단되는 결측값을 제외하고 분석, 이상값을 제거하기 위해 양극단값을 절단하기도 함 극단값 제거 방법보다 극단값 조절 방법을 활용하는 것이 적절하다
    2. 대체법 : 하한,상한값을 결정한 후 이를 기준으로 더 아래,혹은 위일경우 하한,상한으로 대체하는 방법
    3. 변환법 : 극단적인 값으로 인해 이상값이 발생했다면, 자연로그를 취해서 값을 감소
11. 박스플롯을 이용한 이상값 제거
    1. 박스플롯 구성요소 : 1~3사분위수, 최소,최대값, 하위경계(Q1-1.5IQR),최소값(하위경계 최근사값),최대값(상위경계의 최근사값),상위경계(Q#\*1.5IQR), 수염(Q1,Q3 로부터 IQR의 1.5배 내에 있는 가장 멀리 떨어진 데이터까지 이은선) 이상값(수염보다 밖에 있는 값)