# 데이터 분석 개요

9회 기출

**01.** 데이터가 가지고 있는 특성을 파악하기 위해 해당 변수의 분포 등을 시각화하여 분석하는 분석 방식은 무엇인가?

① 전처리분석
② 탐색적자료분석(EDA)
③ 공간분석
④ 다변량분석

**02.** 데이터 마이닝의 모델링에 대한 설명이다. 설명이 가장 잘못된 것은?

① 데이터마이닝 모델링은 통계적 모델링이 아니므로 지나치게 통계적 가설이나 유의성에 집착하지 말아야 한다.
② 모델링 방법은 여러 가지가 있으므로 모델링 시 반드시 다양한 옵션을 줘서 모델링을 수행하여 최고의 성과를 도출하여야 한다.
③ 분석데이터를 학습 및 테스트 데이터로 6:4, 7:3, 8:2 비율로 상황에 맞게 실시한다.
④ 성능에 집착하면 분석 모델링의 주목적인 실무 적용에 반하여 시간을 낭비할 수 있으므로 훈련 및 테스트 성능에 큰 편차가 없고 예상 성능을 만족하면 중단한다.

10회 기출

**03.** 모델링 성능을 평가함에 있어, 데이터마이닝에서 활용하는 평가 기준이 아닌 것은?

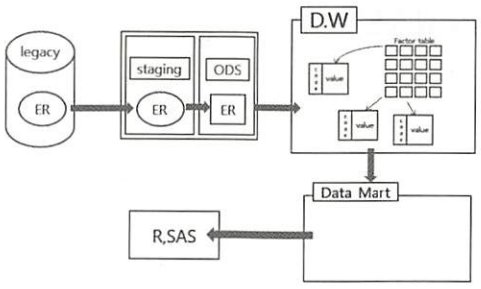
① 정확도(Accuracy)
② 리프트(Lift)
③ 디텍트 레이트(Detect Rate)
④ Throughput

14회 기출

**04.** 탐색적 데이터 분석의 목적은 데이터를 이해하는 것이다. 다음 중 이에 대한 설명으로 가장 부적절한 것은?

① 데이터에 대한 전반적인 이해를 통해 분석 가능한 데이터인지 확인하는 단계이다.
② 탐색적 데이터 분석 과정은 데이터에 포함된 변수의 유형이 어떻게 되는지를 찾아가는 과정이다.
③ 데이터를 시각화하는 것만으로는 이상점(outlier) 식별이 잘 되지 않는다.
④ 알고리즘이 학습을 얼마나 잘 하느냐 하는 것은 전적으로 데이터의 품질과 데이터에 담긴 정보량에 달려 있다.

**05.** 아래의 그림은 데이터 처리 구조를 나타내고 있다. 그림에 대한 설명으로 잘못 된 것은?




① 데이터를 분석에 활용하기 위해 데이터웨어하우스와 데이터마트에서 데이터를 가져 온다.
② 신규시스템이나 DW에 포함되지 않은 데이터는 기존 운영시스템(legacy)에서 직접 데이터를 DW와 전처리 없이 바로 결합하면 된다.
③ ODS는 운영데이터저장소로 기존 운영시스템의 데이터가 정제된 데이터이므로 DW나 DM과 결합하여 분석에 활용할 수 있다.
④ 스테이지 영역에서 가져온 데이터는 정제되어 있지 않기 때문에 데이터의 전처리를 해서 DW나 DM과 결합하여 사용한다.

**06.** 최근 시각화 기법의 활용이 높아지면서 데이터의 특성을 파악하는데 많은 기여를 하고 있다. 다음 중 최근의 시각화의 발전된 형태가 아닌 것은?

① 텍스트 마이닝에서의 워드 클라우드를 통한 그래프화
② SNA(social network analysis)에서 집단의 특성과 관계를 그래프화
③ 통계소프트웨어의 기초통계정보를 엑셀에서 그래프화
④ polygon, heatmap, mosaic graph 등의 그래프 작업

**07.** 대표적인 고급분석으로 데이터에 있는 패턴을 파악해 예측하는 분석으로 데이터가 크고 정보가 다양할수록 보다 활용하기 유리한 분석은 무엇인가?

① 시뮬레이션 ② 통계분석 ③ 데이터 마이닝 ④ 시각화

**08.** 모집단으로부터 추출된 표본의 표본통계량으로부터 모집단의 특성인 모수에 관해 통계적으로 추론하는 통계를 무엇이라고 하는가?

① 가공 통계 ② 기술 통계 ③ 통계분석 ④ 추론 통계

**09.** EDA의 4가지 주제 중 틀린 것은?

① 종속변수 계산
② 저항성의 강조
③ 자료변수의 재표현
④ 그래프를 통한 현시성

**10.** 공간적 차원과 관련된 속성들을 시각화에 추가하여 지도 위에 관련 속성들을 생성하고 크기, 모양, 선 굵기 등으로 구분하여 인사이트를 얻는 분석방법은 무엇인가?

(  )

# R 프로그래밍 기초

**01.** R에 대한 설명으로 옳지 않은 것을 고르시오.

① 뉴질랜드 오클랜드 대학의 로스 이하카와 로버트 젠틀맨에 의해 시작되었다.
② R은 GPL (General Public License)하에 배포되는 S 프로그래밍 언어의 구현으로 GNU S라고도 한다.
③ R은 통계 소프트웨어 개발과 자료 분석에 널리 사용되고 그래픽 처리 기능이 탁월한 언어이다.
④ R의 GNU 일반 공중 사용 허가서(GNU General Public License, GNU GPL 또는 GPL)는 자유 소프트웨어 재단에서 만든 자유 소프트웨어 라이선스이며 리눅스 커널이 이용하는 사용 허가와 동일함으로 리눅스 기반의 언어이다.

8회 기출

**02.** R의 장점으로 옳지 않은 것을 고르시오.

① 오픈 소스이므로 사용자들이 만든 다양한 패키지들을 공유하여 사용 가능하므로 최신 알고리즘을 패키지를 통해 활용하기 쉽다.
② R은 사용자들이 많기 때문에 문제가 발생할 경우, 다양한 사용자들을 통해 문제를 해결하므로 다른 통계패키지에 비해 유지보수가 신속하게 이루어진다.
③ 함수형 언어이기 때문에 다양한 프로그램을 통해 자동화 할 수 있다.
④ 무료로 이용할 수 있다.

**03.** R은 함수형 언어이다. 함수형 언어에 대한 특징으로 옳지 않은 것은?

① 기존에 사용한 함수들을 활용하여 프로그래밍 함으로 프로그램이 더욱 깔끔하고 단축된 코드를 만들 수 있다.
② 함수들을 많이 활용하게 되므로 코드에 대한 수행속도가 늦은 단점이 있다.
③ 함수들을 활용하여 프로그래밍 함으로 코드를 단순화 할 수 있고 디버깅이 쉽다.
④ 병렬 프로그래밍으로 전환이 다른 프로그래밍 언어에 비해 용이하다.

7회 기출

**04.** 다음 중 나머지 세 개의 명령과 결과가 다른 것은?

① z=c(1:3, NA)
is.na(z)
② z<-c(1:3, NA)
is.na(z)
③ z= c(1:3, NA)
z==NA
④ c(1,1,1,2) ==2

7회 기출

**05.** 아래의 R 프로그래밍을 통해 객체 a 에 할당되는 모드가 다른 것을 고르시오.

① a<-c("Tom", "Yoon", "Kim")
② a<-c(pi, "pi", 3.14)
③ a<-c(3.14, pi, TRUE)
④ a<-c("A","B","A","A","B")

11회 기출

**06.** 다음 중 결과가 다른 R코드는?

① a<-seq(1,10,1)
② b<-c(1,10)
③ c<-1:10
④ d<-seq(10,100,10)/10

11회 기출

**07.** 다음 중 아래의 R코드를 수행한 결과에 대한 설명으로 옳은 것은?

```
> c(2, 4, 6, 8) + c(1, 3, 5, 7, 9)
```

① 경고 메시지와 함께 결과가 출력된다.
② 4개의 숫자로 이루어진 벡터가 출력된다.
③ 9개의 숫자로 이루어진 벡터가 출력된다.
④ 에러 메시지가 출력되고, 명령 수행이 중단된다.

**08.** R에서의 데이터 구조에 대한 설명 중에서 잘못된 설명을 고르시오.

① 단일값(scalars)는 원소가 하나인 벡터로 인식 처리하여 R프로그램에서 length(pi)의 결과는 1로 출력된다.
② 행렬(matrices)은 2차원의 벡터를 의미하며 R프로그램에서는 dim()을 활용하여 행렬의 구조를 정의할 수 있다.
③ 요인(factors)은 벡터처럼 생겼지만 원소들이 수준(level)으로 이루어져 있으며, 요인에는 주로 연속형 변수와 집단 분류로 많이 사용된다.
④ 행렬에서 3차원 이상 또는 n차원 이상으로 확대된 형태를 배열(arrays) 이라고 하며 dim( )을 활용하여 배열의 구조를 정의 할 수 있다.

**09.** 아래의 R 프로그램 중 리스트의 원소를 선택하는 방법이 아닌 것은 어느 것인가?

① a<-alist[[2]]
② a<-alist[["name"]]
③ a<-alist[2]
④ a<-alist$name

10회 기출

**10.** 아래의 R코드가 의미하는 것은?

```
> mean(x, na.rm=T)
```

① 이상값을 제외한 X의 평균
② 결측값을 제외한 X의 평균
③ 이상값을 포함한 X의 평균
④ 결측값을 포함한 X의 평균

**11.** 다음 R 함수 중 열의 이름을 붙이는 함수는 무엇인가?

① culname ② culnames ③ colname ④ colnames

10회 기출

**12.** 아래 R코드를 수행한 결과로 적절한 것은?

```
> "+"(2,3)
```

① 에러 메시지가 출력된다.
② 경고 메시지가 출력된다.
③ 숫자 5가 출력된다.
④ 두 개의 원소로 이루어진 벡터가 출력된다.

**13.** 파일에서 데이터를 읽어 들여 가공한 다음, 파일로 저장하는 등 모든 측면에서 매우 직관적이고, 특히 sqldf를 이용할 때 RDBMS의 table 또는 엑셀의 피벗처럼 사용할 수 있는 테이블은 무엇인가?

① list ② matrix ③ vector ④ data.frame

9회 기출

**14.** R에서 결측값을 가르키는 것으로 가장 적절한 것은?

① Inf ② NaN ③ NA ④ dim

15회 기출

**15.** Carseats 데이터프레임은 400개 상점에서 판매 중인 유아용 카시트의 재료이고, Sales 변수는 해당 상점에서 판매된 카시트의 수를 나타낸다. 다음 중 R 패키지에서 Sales 변수의 표준편차를 계산하기 위한 식으로 가장 부적절한 것은?

① stdev(Carseats$Sales)
② sd(Carseats$Sales)
③ sqrt(var(Carseats$Sales))
④ var(Carseats$Sales)^(1/2)

9회 기출

**16.** 다음 중 아래 R 코드의 결과로 적절한 것은?

```
> s<-c("Monday", "Tuesday", "Wednesday")
> substr(s,1,2)
```

① "Mo", "Tu", "We"
② "Monday" "Tuesday"
③ "Mo" "Tu"
④ "Monday"

**17.** 아래 그림과 같이 두개의 데이터 프레임 dfm1, dfm2 를 T_name 이라는 변수로 결합하 고자 하고자 할 때, 사용되는 함수는 어느 것인가?

| T_name | x | y |
|---|---|---|
| T1 | 1.4 | 3.2 |
| T2 | 1.8 | 3.4 |
| T3 | 1.5 | 3.9 |
| T4 | 1.4 | 3.2 |
| T5 | 1.6 | 3.4 |
| T6 | 1.5 | 3.9 |

+

| T_name | z |
|---|---|
| T1 | 5.7 |
| T3 | 5.8 |
| T5 | 6.9 |

=

| T_name | x | y | z |
|---|---|---|---|
| T1 | 1.4 | 3.2 | 5.7 |
| T3 | 1.5 | 3.9 | 5.8 |
| T5 | 1.6 | 3.4 | 6.9 |

① cbind(dfm1, dfm2, by="T_name")
② rbind(dfm1, dfm2, by="T_name")
③ merge(dfm1,dfm2, by="T_name")
④ subset(dfm1,dfm2,by ="T_name")

12회 기출

**18.** 아래 프로그램의 실행 결과로 다음 중 적절한 것은 무엇인가?

```
calculate<-function(a) {
  y=1
  for(i in 1:a) {
    y=y*i
  }
  print(y)
}


calculate(4)
```

① 24 ② 20 ③ 12 ④ 6

**19.** Cars93이라는 데이터프레임에 MPG.city(도심에서의 연비)라는 변수와 Origin(생산지)이라는 변수가 있다고 할 때, 데이터프레임을 생산지 별로 나누려고 한다. R 프로그램으로 적절한 것은?

① split(Cars93, "Origin")
② split("Cars93", "Origin")
③ split(Cars93$MPG.city, Cars93$Origin)
④ split(Cars93, by = Origin)

**20.** 아래와 같은 행렬이 있을 때, 모든 행에 합을 구하기 위한 R 프로그램 중 적절한 것은?

```
> dim(m1)<-c(4,5)
> m1
     [,1] [,2] [,3] [,4] [,5]
[1,] 82.5 79.2 89.5 85.6 80.9
[2,] 89.9 88.2 81.5 91.5 87.2
[3,] 81.9 70.3 89.2 83.2 78.9
[4,] 88.2 83.5 79.8 87.5 82.5
```

① apply(m1, 1, sum)
② apply(m1, 2, sum)
③ lapply(m1, sum)
④ sapply(m2,sum)

**21.** Cars93이라는 데이터프레임에 MPG.city(도심에서의 연비)라는 변수와 Origin(생산지)이라는 변수가 있다고 할 때, 생산지별로 MPG.city의 평균을 구하고자 한다. R 프로그래밍으로 적절한 것은?

① apply(Cars93$MPG.city, Cars93%Origin, mean)
② lapply(Cars93, Cars93%Origin, mean)
③ sapply(Cars93, Cars93%Origin, mean)
④ tapply(Cars93$MPG.city, Cars93%Origin, mean)

5회 기출

**22.** 단어나 문장에 포함되어 있는 문자열의 길이를 구하고자 할 때, R 프로그램으로 적절한 것은?

① nchar("statistics")
② length("statistics")
③ substr("statistics")
④ paste ("statistics")

**23.** 문자열 "statistics"에서 "at"를 추출하고자 할 때, R 프로그램으로 적절한 것은?

① substr("statistics", 3, 2)
② substr("statistics", 3, 4)
③ strisplit("statistics", 3, 2)
④ strsplit ("statistics", 3, 4)

**24.** R을 GUI 환경에서 보다 편리하게 사용할 수 있도록 도와주는 패키지는 무엇인가?

① R stdio ② rattle ③ shiny ④ impala

10회 기출

**25.** 아래 R 코드의 출력 결과는?

```
> f <- function(x,a) return((x-a)^2)
> f(1:2,3)
```

(  )

8회 기출

**26.** R에서 다음의 명령을 수행했을 때 출력되는 결과는?

```
x<-c(1,2,3,NA)
mean(x)
```

(  )

7회 기출

27. 출력결과는?

```
x<-1:100
sum(x>50)
```

(  )

28. A반과 B반 학생들이 동일한 과목을 들었다고 하자. A반과 B반 학생 모두를 대상으로 과목 별 성적의 평균을 구하려고 할 때, A반 학생 데이터와 B반 학생 데이터를 class 라는 변수를 기준으로 합치려고 한다. R로 프로그램을 작성하시오.

(  )

29. 아래의 표와 같이 여러 학과 학생들의 과목별 성적을 데이터 프레임으로 구성하였다. 데이터 프레임명은 test 라고 할 때, 경영학과 학생들의 데이터만 조회하고자 한다. R로 프로그래밍 하시오.

| 학과 | 학년 | 성별 | 이름 | 실용컴퓨터 | 영어회화 | 한문 | 총점 |
|---|---|---|---|---|---|---|---|
| 경영학과 | 1 | 여 | 김지영 | 85 | 75 | 86 | 246 |
| 경영학과 | 1 | 여 | 이소연 | 75 | 65 | 78 | 218 |
| 경영학과 | 1 | 남 | 이진혁 | 96 | 77 | 67 | 240 |
| 데이터정보학과 | 3 | 남 | 김영수 | 45 | 78 | 56 | 179 |
| 데이터정보학과 | 1 | 남 | 김민수 | 86 | 87 | 84 | 257 |
| 데이터정보학과 | 1 | 여 | 박미혜 | 100 | 92 | 96 | 288 |
| 데이터정보학과 | 1 | 남 | 최성호 | 87 | 95 | 92 | 274 |
| 영문학과 | 4 | 여 | 김동수 | 68 | 75 | 78 | 221 |
| 영문학과 | 2 | 남 | 이민지 | 99 | 86 | 86 | 271 |

(  )

30. SQL을 활용하거나 SAS에서 porc sql로 작업하던 사용자들에게 R 프로그램에서 지원해 주는 패키지는 무엇인가?

(  )

# 데이터 마트

6회 기출

**01.** 데이터 웨어하우스와 사용자의 중간층에 위치한 것으로, 하나의 주제 또는 하나의 부서 중심의 데이터 웨어하우스라고 할 수 있는 데이터베이스는 무엇인가?

① 데이터마트 ② 모델링 ③ 관계형 데이터베이스 ④ 빅데이터

**02.** 변수를 조합해 변수명을 만들고 변수들을 시간, 상품 등의 차원에 결합해 다양한 요약변수와 파생변수를 쉽게 생성하여 데이터 마트를 구성할 수 있는 패키지는 무엇인가?

① ETL ② reshape ③ OLAP ④ rattle

10회 기출

**03.** 파생변수는 사용자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수이다. 다음 중 파생변수의 설명으로 적절한 것은?

① 파생변수는 매우 주관적인 변수일 수 있으므로 논리적 타당성을 갖춰야 한다.
② 파생변수는 많은 모델에서 공통적으로 많이 사용될 수 있다.
③ 파생변수는 재활용성이 높다.
④ 파생변수는 다양한 모델을 개발해야 하는 경우, 효율적으로 사용할 수 있다.

10회 기출

**04.** 많은 기업에서 평균거래주기를 3~4배 이상 초과하거나 다음 달에 거래가 없을 것으로 예상되는 고객을 ( ㉠ )으로 정의하고 있다. 다음 중 ( ㉠ )에 가장 적절한 것은?

① 신규고객 ② 우량고객 ③ 가망고객 ④ 휴면고객

6회 기출

**05.** 아래 표는 데이터의 변경을 통해 새로운 구조의 데이터셋을 구성하고자 할때 사용하는 R 프로그램 중 melt함수와 cast 함수의 예시이다. 데이터셋 MD를 새로운 데이터 형태로 변경하기 위한 cast 함수를 활용한 R 프로그램 중 옳은 것은?

〈DATA 명 : MD〉

| ID | Time | Variable | Value |
|---|---|---|---|
| 1 | 1 | X1 | 5 |
| 1 | 2 | X1 | 3 |
| 2 | 1 | X1 | 6 |
| 2 | 2 | X1 | 2 |
| 1 | 1 | X2 | 6 |
| 1 | 2 | X2 | 5 |
| 2 | 1 | X2 | 1 |
| 2 | 2 | X2 | 4 |

〈새로운 데이터〉

| ID | Variable | Time1 | Time2 |
|---|---|---|---|
| 1 | X1 | 5 | 3 |
| 1 | X2 | 6 | 5 |
| 2 | X1 | 6 | 2 |
| 2 | X2 | 1 | 4 |

① cast(md, id~variable +time)

② cast(md, id+variable~time)

③ cast(md, id+time~variable)

④ cast(md, id~variable, mean)

9회 기출

**06.** 아래의 정의가 가리키는 데이터 마트의 구성요소로 가장 적절한 것은?

> 특정한 의미를 갖는 작위적 정의에 의한 변수로, 사용자가 특정 조건을 만족하거나 특정 함수에 의해 값을 만들어 의미를 부여한 변수

① 반응변수 ② 파생변수 ③ 설명변수 ④ 요약변수

7회 기출

**07.** 아래의 왼쪽 자료를 오른쪽의 형태로 변환하기 위한 명령어로 적절한 것은?

>head(airquality, 10)

| | Ozone | Solar.R | Wind | Temp | Month | Day |
|---|---|---|---|---|---|---|
| 1 | 41 | 190 | 7.4 | 67 | 5 | 1 |
| 2 | 36 | 118 | 8 | 72 | 5 | 2 |
| 3 | 12 | 149 | 12.6 | 74 | 5 | 3 |
| 4 | 18 | 313 | 11.3 | 62 | 5 | 4 |
| 5 | NA | NA | 14.3 | 56 | 5 | 5 |
| 6 | 28 | NA | 14.9 | 66 | 5 | 6 |
| 7 | 23 | 299 | 8.6 | 65 | 5 | 7 |
| 8 | 19 | 99 | 13.8 | 59 | 5 | 8 |
| 9 | 8 | 19 | 20.1 | 61 | 5 | 9 |
| 10 | NA | 194 | 8.6 | 69 | 5 | 10 |

> aqm

| | month | day | variable | value |
|---|---|---|---|---|
| 1 | 5 | 1 | Ozone | 41 |
| 2 | 5 | 2 | Ozone | 36 |
| 3 | 5 | 3 | Ozone | 12 |
| 4 | 5 | 4 | Ozone | 18 |
| 5 | 5 | 6 | Ozone | 28 |
| 6 | 5 | 7 | Ozone | 23 |
| 7 | 5 | 8 | Ozone | 19 |
| | | | | |
| 115 | 9 | 29 | Ozone | 18 |
| 116 | 9 | 30 | Ozone | 20 |
| 117 | 5 | 1 | Solar.R | 190 |
| 118 | 5 | 2 | Solar.R | 118 |
| 119 | 5 | 3 | Solar.R | 149 |
| | | | | |
| 564 | 9 | 28 | Temp | 75 |
| 565 | 9 | 29 | Temp | 76 |
| 566 | 9 | 30 | Temp | 68 |

① aqm<-melt(airquality, id=c("Month","Day"), na.rm=TRUE)

② aqm<-melt(airquality, id=c("Month","Day"))

③ aqm<-melt(airquality, id=c("Ozone","Solar.R","Wind","Temp"), na.rm=TRUE)

④ aqm<-melt(airquality, id=c ("Ozone","Solar.R","Wind","Temp"))

**08.** "iris"라는 데이터셋에서 데이터의 내용을 조회할 때, R프로그램으로 적절한 것은?

① plyr("select*from iris")
② sql("select*from iris")
③ mysql("select*from iris")
④ sqldf("select*from iris")

14회 기출

**09.** chickwts 데이터프레임은 여섯가지 종류의 닭 사료 첨가물(feed)과 각 사료를 먹인 닭의 무게(weight)를 변수로 가진다. 아래의 (1)의 기초통계량과 각 feed별 weight의 평균을 계산하여, 아래 (2)와 같은 결과물을 만들기 위한 코드로 다음 중 가장 적절한 것은?

(1)

```
> head(chickwts)
  weight      feed
1    179 horsebean
2    160 horsebean
3    136 horsebean
4    227 horsebean
5    217 horsebean
6    168 horsebean
> summary(chickwts)
     weight             feed
 Min.   :108.0   casein   :12
 1st Qu.:204.5   horsebean:10
 Median :258.0   linseed  :12
 Mean   :261.3   meatmeal :11
 3rd Qu.:323.5   soybean  :14
 Max.   :423.0   sunflower:12
```

(2)

```
       feed groupmean
1    casein  323.5833
2 horsebean  160.2000
3   linseed  218.7500
4  meatmeal  276.9091
5   soybean  246.4286
6 sunflower  328.9167
```

① ddply(chickwts, ~feed, groupmean=mean(weight))
② ddply(chickwts, weight~feed, summarize, groupmean=mean(weight))
③ ddply(chickwts, ~feed, summarize, groupmean=mean(weight))
④ ddply(chickwts, weight~feed, groupmean=mean(weight))

7회 기출

**10.** 다음 중 결측치에 대한 설명으로 가장 부적절한 것은?

① 해당 칸이 비어있는 경우 결측치 여부는 알기 쉽다.
② 관측치가 있지만 실상은 default 값이 기록된 경우에도 결측치로 처리해야 하는 것이 바람직하다.
③ 결측치가 있는 경우 다양한 대치(Imputation)방법을 사용하여 완전한 자료로 만든 후 분석을 진행할 수 있다.
④ 결측치가 20% 이상인 경우에는 해당 변수를 제거하고 분석해야 한다.

**11.** 다음은 결측값을 확인하고 결측값을 대치하는데 활용되는 R 함수들이다. 설명이 잘못된 것을 고르시오.

① complete.cases( ) : 데이터 내 레코드에 결측값이 있으면 TRUE, 없으면 FALSE를 반환하는 함수
② is.na( ) : 결측값이 NA인지 여부를 판단하여 반환하는 함수
③ knnImputation( ) : NA 값을 k 최근 이웃 분류 알고리즘을 사용하여 대치하는 함수로 k개 주변 이웃까지의 거리를 고려하여 가중 평균한 값을 대치해 주는 함수
④ rfImpute( ) : 랜덤 포레스트 모형의 경우, 결측값이 있으면 에러를 발생하기 때문에 랜덤포레스트 패키지에서 NA 결측값을 대치하도록 하는 함수

**12.** 데이터프레임과 유사하지만 보다 빠른 그룹핑과 ordering, 짧은 문장 지원 측면에서 데이터 프레임보다 매력적으로 활용이 가능한 패키지는 무엇인가?

① sqldf ② OLAP ③ data.table ④ plyr

14회 기출

**13.** 이상치를 찾는 것은 데이터 분석에서 데이터 전처리를 어떻게 할지 검정할 때 사용할 수 있다. 다음 중 상자그림을 이용하여 이상치를 판정하는 방법에 대한 설명으로 가장 부적절한 것은?

① IQR=Q3-Q1이라고 할 때, Q1-1.5*IQR$<x<$Q3+1.5*IQR을 벗어나는 $x$를 이상치라고 규정한다.
② 평균으로부터 3*표준편차 벗어나는 것들을 비정상이라 규정하고 제거한다.
③ 이상치는 변수의 분포에서 벗어난 값으로 상자 그림을 통해 확인할 수 있다.
④ 이상치는 분포를 왜곡할 수 있으나 실제 오류인자에 대해서는 통계적으로 실행하지 못하기 때문에 제거여부는 실무자들을 통해서 결정하는 것이 바람직하다.

11회 기출

**14.** 다음 중 이상값 검색을 활용한 응용시스템으로 가장 적절한 것은?

① 장바구니분석 시스템
② 데이터 마트
③ 교차판매 시스템
④ 부정사용방지 시스템

5회 기출

**15.** 이상치에 대한 설명으로 가장 부적절한 것은?

① 군집분석을 이용하여 다른 데이터들과 거리상 멀리 떨어진 데이터를 이상치로 판정한다.
② 데이터를 측정과정이나 입력하는 과정에서 잘못 포함된 이상치는 삭제한 후 분석한다.
③ 설명변수의 관측치에 비해 종속변수의 값이 상이한 값을 이상치라 한다.
④ 통상 평균으로부터 표준편차의 3배가 되는 점을 기준으로 이상치를 정의한다.

**16.** 다음은 이상값(outlier)에 대한 설명이다. 잘못 설명한 내용을 고르시오.

① 부정사용방지 시스템이나 부도예측시스템에서는 이상값(outlier)이라도 의미가 있으므로 제거하지 않는다.
② 이상값 인식에 있어서 가장 많이 활용하는 방법은 ESD(Extreme Studentized Deviation)으로 평균에서 3 표준편차를 벗어나는 경우 이상값으로 인식하는 방법이다.
③ 이상값의 처리에 있어서 극단값 절단 방법과 조정 방법이 있으며 조정의 경우, 제거 방법에 비해 데이터 손실율이 높아 설명력이 낮아지는 단점이 있다.
④ 의도하지 않게 잘못 입력된 데이터인 경우 bad data 에 해당되며 이러한 경우, 데이터를 제거하여 분석한다.

**17.** 결측치(Missing data) 핸들링은 데이터분석을 위한 전처리 작업에서 가장 중요한 단계 중에 하나이다. R 프로그램에서 결측치의 표현으로 맞는 것은?

① Missing
② 999999999
③ NaN(Not a Number)
④ NA(Not Available)

**18.** 데이터 전처리 단계에서 데이터의 이상치(Outlier) 에 대한 설명으로 틀린 것은?

① 최대값과 최소값
② 데이터 입력 시 오타로 인해 잘못 입력된 경우
③ 분석 목적에 부합되지 않아 제거해야 하는 경우
④ 부정사용방지 시스템에서 의도된 이상 값

9회 기출

**19.** 아래는 이상치(Outlier) 탐지에 대한 설명이다. 다음 중 이상치를 유용하게 사용하는 분야의 예로 부적절한 것은?

> 이상치(Outlier) 탐지의 목적은 대부분의 객체들과 다른 객체들을 찾는 것이다. 이상치 탐지는 속성값들의 일반적인 값들과 상당히 편차가 큰 값을 가지므로 편차 탐지(deviation detection)라고도 한다. 그러나 이상치는 반드시 비정상적인 객체를 의미하지는 않는다.

① 사기탐지 - 도난당한 신용카드의 구매 행위는 원 소유자의 행위와 다를 수 있다. 평상시의 행위와 다른 구매패턴을 조사하여 사기를 탐지할 수 있다.

② 환경파괴 - 자연 세계에서는 환경에 중요한 영향을 줄 수 있는 홍수, 가뭄 같은 사건들이 있다. 그러나 이러한 사건은 정상적인 환경에서 발생하는 사건으로 해석할 수 있다.

③ 의료 - 특정 환자에게 보이는 예외적인 증세나 검사 결과는 잠재적인 건강 문제를 나타낸다.

④ 침입탐지 - 컴퓨터 네트워크에 대한 공격은 보편화되었다. 침입의 다수는 네트워크에 대한 예외적인 행위를 감시하는 경우에 탐지할 수 있다.

11회 기출

**20.** 평균으로부터 t standard deviation 이상 떨어져 있는 값들을 이상값(outlier)으로 판단하고 t는 3으로 설정하는 이상값 검색 알고리즘은?

(  )