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?

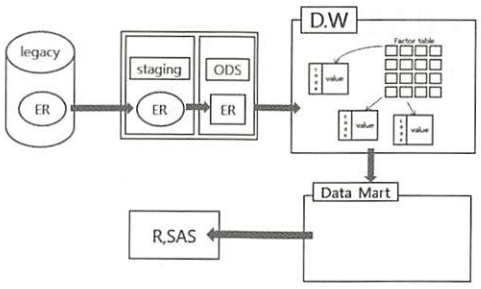
① 정확도(Accuracy)
② 리프트(Lift)
③ 디텍트 레이트(Detect Rate)
④ Throughput

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?

( )

정답 및 해설

01	②
02	②
03	④
04	③
05	②
06	③
07	③
08	④
09	①
10	공간분석(spacial analysis)

01. EDA는 매우 시간이 많이 필요한 일로 최근에는 EDA를 자동으로 신속하게 수행해 유의미한 값만 파악해 데이터 마트로 만든 후 모델링 업무로 진행하는 게 일반적이다. (**정답 : ②**)

02. 반드시 다양한 옵션을 줘서 모델링을 수행하지 않고, 충분한 시간이 있으면 다양한 옵션을 줘서 시도하는 것이고 일정 성과가 나오면 해석과 활용 단계로 진행할 수 있도록 의사결정 해야 한다. (**정답 : ②**)

03. 데이터 마이닝에서는 정확도, 정밀도, 디텍트 레이트, 리프트 등의 값으로 판단하고 시뮬레이션에서는 Throughput, Average Waiting Time, Average Queue Length, Time in System 등의 지표가 활용된다. (**정답 : ④**)

04. 상자그림(Box Plot)등을 그리면 이상치를 식별하기 쉽다. (**정답 : ③**)

05. 신규 시스템이나 스테이징 영역의 데이터는 정제되지 않았기 때문에 정제하고 DW나 DM과 결합해야 한다. (**정답 : ②**)

06. 엑셀의 그래프는 최근 시각화 기술의 발전된 형태가 아니라 기존에 기술이다. (**정답 : ③**)

07. 대용량 데이터에서 패턴을 파악해서 예측하는 분석 방법은 데이터마이닝 방법이다. (**정답 : ③**)

08. 추론(추측)통계는 모집단으로부터 추출된 표본의 표본통계량으로부터 모집단의 특성인 모수에 관한 통계적으로 추론하는 절차이다. (**정답 : ④**)

09. EDA의 4가지 주제는 저항성의 강조, 잔차 계산, 자료변수의 재표현, 그래프를 통한 현시성이다. (**정답 : ①**)

10. 지도위에 공간과 관계된 속성들을 다양한 표현으로 시각화하는 방법은 공간 분석이다.
정답 : 공간분석(spacial analysis)

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?

( )

정답 및 해설

01	④	11	④	21	④
02	②	12	③	22	①
03	②	13	④	23	②
04	③	14	③	24	②
05	③	15	①	25	4 1
06	②	16	①	26	NA
07	①	17	③	27	50
08	③	18	①	28	merge(A,B,by="class")
09	③	19	③	29	subset(test, subset=(학과==경영학과))
10	②	20	①	30	sqldf()

01. R은 윈도우, 맥, 리눅스 운영체계에서 모두 사용 가능하다. (**정답 : ④**)

02. R은 사용자들이 많기 때문에 문제가 발생할 경우, 다양한 사용자들을 통해 다양한 의견들을 들을 수 있으나 적절한 해결책을 찾기 위해서는 시간과 노력이 필요합니다. SAS나 SPSS와 같은 솔루션의 경우, 문제가 발생할 경우 해당 업체를 통해 유지보수가 신속하게 이루어진다. (**정답 : ②**)

03. 함수형 언어이므로 적절한 함수를 적용하여 프로그래밍 하면 수행 속도도 매우 빠르게 된다. (**정답 : ②**)

04. ①, ②, ④의 결과는 모두 FALSE FALSE FALSE TRUE 이지만, ③의 경우에는 NA NA NA NA가 나타난다. (**정답 : ③**)

05. ①, ②, ④의 결과는 모두 character 이지만, ③의 경우에는 numeric이다. (**정답 : ③**)

06. ①, ③, ④의 결과는 모두 1 2 3 4 5 6 7 8 9 10이지만, ②의 경우에는 1 10이다. (**정답 : ②**)

07. 아래의 R 코드를 실행시키면 '두 객체의 길이가 서로 배수관계에 있지 않습니다'라는 경고 메시지가 뜨고 결과도 출력된다. (**정답 : ①**)

08. 요인은 범주형 변수와 집단 분류에 많이 사용된다. (**정답 : ③**)

09. 보기 ③번은 리스트의 주소를 나타내는 명령어이다. (**정답 : ③**)

10. 해당 R 코드 중 na.rm은 결측치를 제외하느냐에 대한 물음이며, T는 TRUE로서 결측치를 제외하겠다는 의미이다. (**정답 : ②**)

11. colnames( )는 열의 이름을 조회한다. (**정답 : ④**)

12. 아래의 코드를 실행하면 숫자 5가 출력된다. (**정답 : ③**)

13. 데이터프레임은 관찰된 결과로 된 테이블이다. 데이터 프레임은 테이블로 된 사각형의 데이터 구조여서 열과 행이 있다. 하지만 행렬로 구현되지는 않았다. 오히려 리스트라고 볼 수 있다. 그리고 가장 자주 사용되고 편리한 데이터처리 방식이 데이터프레임이다. 특히 sqldf를 이용할 때, RDBMS의 table 또는 엑셀의 피벗처럼 사용할 수 있다. (**정답 : ④**)

14. Inf는 무한대, NaN은 Not a Number, dim은 행렬의 차원을 나타낸다. (**정답 : ③**)

15. R에서 표준편차를 계산하기 위해 사용하는 함수가 아닌 것은 stdev()함수이다. (**정답 : ①**)

16. 아래의 코드를 실행하면 "Mo", "Tu", "We"가 나타난다. (**정답 : ①**)

17. 두 개의 테이블을 하나로 변경할 때 merge 함수를 사용한다. (**정답 : ③**)

18. Calculate(4)를 실행 했을 때, (1). y=1, i=1 → y=1, (2). y=1, i=2 → y=2, (3). y=2, i=3 → y=6, (4). y=6, i=4 → y=24 이므로 24가 출력 (**정답 : ①**)

19. split(분석대상, 분할대상)으로 활용해야 한다. 그러므로 데이터프레임을 생산지별로 나누려면 split(Cars93$MPG.city, Cars93$Origin)을 활용해야 한다. (**정답 : ③**)

20. apply 함수에서 두 번째 인자가 1이면 행, 2이면 열의 자료를 적용한다. (**정답 : ①**)

21. tapply 함수에서 인자에는 정확한 위치의 변수명을 지정하고 적용할 함수를 할당해야 한다. (**정답 : ④**)

22. 단어나 문장에 포함되어 있는 문자열의 길이를 구하고자 할 때는 nchar함수를 사용한다. (**정답 : ①**)

23. substr 함수는 추출하고자 하는 시작위치와 끝위치를 지정한다. (**정답 : ②**)

24. 래틀(rattle)은 R을 GUI 환경에서 보다 편리하게 사용할 수 있도록 도와주는 패키지다. 특히 컴퓨터 언어에 익숙하지 않은 초보자들이 별다른 사전지식 없이도 R을 이용할 수 있게 해준다. (**정답 : ②**)

25. **정답 : 4 1**

26. **정답 : NA**

27. **정답 : 50**

28. **정답 : merge(A, B, by="class")**

29. **정답 : subset(test, subset=(학과==경영학과))**

30. **정답 : sqldf()**