# 데이터 분석 준전문가
# 모의고사(ADsP)

---

| 출제 | 데이터에듀 |
| --- | --- |
| 문항수 | 객관식 : 40<br>주관식 : 10 |

제1회 모의고사

제2회 모의고사

# 데이터 분석 준전문가 모의고사(ADsP) 1회

**출제** 데이터에듀
**문항수** 객관식 : 40 / 주관식 : 10

▶ 배점 안내 : 객관식(40문항) 각 2점 / 주관식(10문항) 각 2점

과목1. 데이터 이해 - 8문항

**01.** 데이터베이스의 특징으로 가장 부적절한 것은?

① 데이터베이스는 여러 사용자가 서로 다른 목적으로 데이터를 공동으로 이용할 수 있도록 구성되어 있다.
② 데이터베이스는 통합된 데이터(Integrated Data)다.
③ 데이터베이스는 변화하는 데이터로 데이터의 삽입, 삭제, 갱신을 한다고 하더라도 항상 현재의 정확한 데이터를 유지해야 한다.
④ 데이터베이스는 검색기능을 가지고 있으므로 다양한 방법으로 필요한 정보를 검색할 수 있다.

**02.** DIKW 피라미드의 계층 중 "B마트 보다 상대적으로 저렴한 A마트에서 연필을 사야겠다."의 내용에 해당하는 계층은 무엇인가?

① 지혜 ② 지식 ③ 정보 ④ 데이터

**03.** 다음 중 데이터 사이언티스트(Data Scientist)에게 요구되는 소프트 역량(Soft Skill)이 아닌 것은?

① 이론적 지식
② 창의적 사고
③ 커뮤니케이션 기술
④ 비쥬얼라이제이션을 활용한 설득력

**04.** 다음 중 빅데이터 분석에 경제성을 제공해 준 결정적인 기술로 가장 적절한 것은?

① 저장장치 비용의 지속적인 하락
② 텍스트 마이닝
③ 클라우드 컴퓨팅
④ 스마트폰의 급속한 확산

**05.** 아래와 같은 SQL 문장을 사용할 때, 출력되는 결과로 옳은 것은?

```
select customer_name 고객명, e_customer_name 고객 영문명
from customer
where e_customer_name like '_A%';
```

① 영문명이 A로 시작하는 고객들의 이름
② 영문명에 A를 포함한 고객들의 비율
③ 위치 상관없이 영문명에 A를 포함하는 고객들의 이름
④ 영문명에 두 번째 문자가 A인 고객들의 이름

**06.** 인터넷 등 각종 경로로 정보를 수집하는 구글은 이미 지난 2010년에 서비스 이용자가 1시간 뒤에 어떤 일을 할지 87% 정확도로 예측할 수 있는 데이터와 분석 신뢰도를 확보하고 있다고 했다. 또, 여행사실을 트위트한 사람의 집을 강도가 노리는 고전적 사례도 발생했다. 이러한 사례를 통해 알 수 있는 빅데이터 시대의 위기 요인으로 적절한 것은?

① 소셜 네트워크 ② 책임 원칙 훼손 ③ 데이터 오용 ④ 사생활 침해

**07.** 사물끼리 정보를 주고 받는 사물인터넷 시대를 빅데이터의 관점에서 바라볼 때 다음 중 사물인터넷의 의미로 가장 적절한 것은?

① 모든 것의 데이터화(Datafication)
② 서비스 지능화(Intelligent Service)
③ 분석 고급화(Advanced Analytics)
④ 정보 공유화(Information Sharing)

**08.** 빅데이터와 데이터 사이언스의 미래를 위한 외부 환경적 측면에서 인문학의 열풍의 원인을 설명한 것 중 옳지 않은 것은?

① 단순세계화에서 복잡한 세계화로 변화하는 과정에서 인문학의 중요성을 인식하여야 한다.
② 비즈니스의 화두가 글로벌 네트워크를 통한 대량공급으로 변함에 따라 가격 인하 정책의 성공을 위해서는 인문학이 중요하다.
③ 비즈니스 중심이 제품생산에서 서비스로 이동함에 따라 인문학의 중요성이 증가하고 있다.
④ 경제와 산업의 논리가 생산에서 시장 창조로 변화하면서 인문학의 중요성이 증가하고 있다.

과목2. 데이터 분석 기획 - 8문항

**09.** 분석은 분석의 대상(What) 및 분석의 방법(How)에 따라 4가지 분석 주제로 나눌 수 있다. 분석의 대상이 명확하게 무엇인지 모르면서 기존 분석 방법으로 새로운 분석을 수행하는 방식의 분석 주제 유형은 무엇인가?

① 최적화(Optimization)
② 통찰(Insight)
③ 솔루션(Solution)
④ 발견(Discovery)



**10.** 다음 중 성공적인 분석을 위해서 고려해야 할 요소로 가장 부적절한 것은?

① 분석 데이터에 대한 고려
② 활용 가능한 유즈케이스 탐색
③ 원점에서 솔루션 탐색
④ 장애 요소에 대한 사전 계획 수립

**11.** 분석 과제를 발굴하기 위한 접근법 중 하향식 접근방법의 과정이 아닌 것은?

① 기업의 내/외부 환경을 포괄하는 비즈니스 모델과 외부 사례를 기반으로 문제를 탐색한다.
② 기업내부의 과거 데이터를 무조건 결합 및 활용한다.
③ 식별된 비즈니스 문제를 데이터의 문제로 변환하여 정의한다.
④ 도출된 분석 문제나 가설에 대한 대안을 과제화하기 위해 타당성을 평가한다.

12. 분석기회 발굴의 범위 중 시장니즈 탐색 관점에서 고객 니즈의 변화에 해당하는 것이 아닌 것은?

① 고객 ② 채널 ③ 영향자들 ④ 대체재

13. 거시적 관점의 메가 트렌드에서 현재의 조직과 해당 산업에 폭넓게 영향을 미치는 사회·경제적 요인인 STEEP로 폭넓게 기회를 탐색한다. STEEP 중 Political(정치영역)의 주요관점에 대한 설명으로 가장 적절한 것은?

① 주요 정책방향, 정세, 지정학적 동향 등 거시적인 흐름을 토대로 분석기회를 도출한다.
② 산업과 경제 구조 변화 동향에 따른 시장의 흐름을 파악하여 분석기회를 도출한다.
③ 정부, 사회단체, 시민사회의 환경에 관한 관심과 규제 동향을 파악하여 분석기회를 도출한다.
④ 과학, 기술, 의학 등 최신 기술의 등장 및 변화를 파악하여 분석기회를 도출한다.

14. 분석 프로젝트 영역별 주요 관리 항목이 아닌 것은?

① 품질 ② 시간 ③ 가격 ④ 자원

15. 다음 중 분석 과제 관리 프로세스에 대한 설명으로 가장 적절하지 않은 것은 무엇인가?

① 분석 아이디어 발굴, 분석과제 후보제안, 분석과제 확정 프로세스는 과제 발굴 단계에 속해 있다.
② 분석과제로 확정되면 분석 과제를 풀(Pool)로 관리한다.
③ 분석과제 중에 발생된 시사점과 분석 결과물은 풀(Pool)로 관리하고 공유된다.
④ 과제 수행 단계에서는 팀 구성, 분석과제 식별, 분석과제 진행관리, 결과 공유 프로세스가 있다.

16. 다음 데이터 분석 조직의 유형 중 별도의 분석 조직이 없고 해당 업무부서에서 분석을 수행하는 방식에 해당하는 것은?

① 기능형 ② 분산형 ③ 복합형 ④ 집중형

과목3. 데이터 분석 - 24문항

17. 모형을 개발하여 운영상황에서 실제 테스트를 할 때 모형 개발 데이터를 통해서는 높은 적중률을 보이지만 테스트 데이터에서는 적중률이 떨어져 적중률을 유지하지 못하는 것을 무엇이라고 하는가?

① 일반화 ② 과대적합 ③ 미적합 ④ 과소평가

**18.** 측정대상이 갖고 있는 속성의 양을 측정하는 것으로 측정결과가 숫자로 표현되나 해당 속성이 전혀 없는 상태인 절대적인 영점이 없어 두 관측 값 사이의 비율은 별 의미가 없게 된다. 온도, 지수 등이 해당되는 이 척도는 무엇인가?

① 명목척도 ② 순서척도 ③ 구간척도 ④ 비율척도

**19.** 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.

**20.** 다음 중 모분산의 추론에 대한 설명으로 적절하지 않은 것은 무엇인가?

① 이표본에 의한 분산비 검정은 두 표본의 분산이 동일한지를 비교하는 검정으로 검정통계량은 F분포를 따른다.
② 모분산이 추론의 대상이 되는 경우는 모집단의 변동성 또는 퍼짐의 정도에 관심이 있을 때이다.
③ 모집단이 정규분포를 따르지 않더라도 중심극한 정리를 통해 정규 모집단으로부터의 모분산에 대한 검정을 유사하게 시행할 수 있다.
④ 평균모집단에서 n개를 단순임의 추출한 표본의 분산은 자유도가 n-1 인 t 분포를 따른다.

**21.** 다음 다중회귀분석을 위해 사용되는 변수선택방법에 대한 설명 중 변수선택방법과 설명이 잘못 연결되어 있는 것은?

① 전진선택법(Forward Selection)은 상수항만 포함된 모형에서 출발하여 설명력이 좋은 변수를 하나씩 추가하는 방법이다.
② 단계적 방법(Stepwise Method)은 설명력이 나쁜 변수를 제거하거나 모형에서 제외된 변수 중 모형의 설명력을 가장 잘 개선하는 변수를 추가하는 방법이다.
③ 후진제거법(Backward Elimination)은 모든 변수가 포함된 모형에서 설명력이 나쁜 변수를 하나씩 제거하는 방법이다.
④ 최적선택법(Optimum Selection)은 전진선택법과 후진제거법을 결합한 방법으로 회귀식이 최적의 변수를 선택하도록 하는 방법이다.

**22.** 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-Q1 이라고 할 때, $Q1-1.5*IQR < x < Q3+1.5*IQR$ 을 벗어나는 x를 이상치라고 규정한다.

② 평균으로부터 3*표준편차 범위를 벗어나는 것들을 비정상이라 규정하고 제거한다.

③ 이상치는 변수의 분포에서 벗어난 값으로 상자 그림을 통해 확인할 수 있다.

④ 이상치는 분포를 왜곡할 수 있으나 실제 오류 인자인지에 대해서는 통계적으로 판단하지 못하기 때문에 제거여부는 실무자들을 통해서 결정하는 것이 바람직하다.

**23.** 통계분석에서 자료를 수집하고 그 수집된 자료로부터 어떤 정보를 얻고자 하는 경우에는 항상 수집된 자료가 특정한 확률분포를 따른다고 가정한다. 다음 중 연속형 확률분포가 아닌 것은?

① 이항분포 ② 정규분포 ③ t분포 ④ F분포

**24.** 다음 표본 추출 방법에 관한 설명 중 잘못된 것은 무엇인가?

① 표본의 크기를 결정할 때 가장 중요한 부분은 표본이 모집단을 얼마나 설명하는지에 대한 대표성의 확보이다.

② 단순랜덤추출법은 모집단에서 샘플을 뽑을 때 각각의 샘플이 모두 동등한 확률을 가지고 무작위로 추출되는 방법이다.

③ 계통추출법은 모집단을 군집으로 구분하고 선정된 군집의 원소를 모두 샘플로 추출하는 다단계 추출 방법이다.

④ 층화추출법은 모집단을 몇 개의 집단으로 구분하고, 각 집단의 크기와 분산을 고려하여 각 집단마다 샘플을 추출하는 방법이다.

**25.** 다음 중 비모수검정이 아닌 것을 고르시오.

① 윌콕슨의 순위합 검증

② 맨-휘트니의 U검정

③ 스피어만의 순위상관계수

④ 자기상관검정

**26.** 두 변량 X, Y의 상관분석에 관한 내용이다. 설명이 옳지 않은 것은?

① 등간척도로 측정된 두 변수 간의 상관관계는 피어슨 상관계수(Pearson Correlation)를 통해 확인할 수 있다.

② 상관계수가 0이면 두 변량 X, Y사이에 선형관계가 없다.

③ 서열척도로 측정된 두 변수간의 상관관계는 스피어만 상관계수(Spearman Correlation)를 통해 확인할 수 있다.

④ R에서 상관계수를 구하기 위해서는 rcor( )함수를 사용하면 되고 type인자를 통해 피어슨과 스피어만 상관계수를 선택할 수 있다.

**27.** 다음 중 회귀분석에서 나온 결정계수( $R^2$)에 대한 설명으로 옳지 않은 것은?

① 총제곱의 합 중 설명된 제곱의 합의 비율을 뜻한다.

② 종속변수에 미치는 영향이 적은 독립변수가 추가된다면 결정계수는 변하지 않는다.

③ $R^2$의 값이 클수록 회귀선으로 실제 관찰치를 예측하는 데 정확성이 높아진다.

④ 독립변수와 종속변수 간의 표본상관계수 r의 제곱값과 같다.

28 DAY

**28.** 다음 시계열 분석의 기초가 되는 개념인 정상성(Stationarity)의 특징에 관한 설명이다. 설명이 옳지 않은 것은?

① 평균이 일정하다. 즉 모든 시점에 대한 일정한 평균을 가진다.

② 시계열 분석에서 비정상 시계열 자료는 시계열 분석을 할 수 없다.

③ 분산도 시점에 의존하지 않는다.

④ 공분산은 단지 시차에만 의존하고 실제 어느 시점 t, s에는 의존하지 않는다.

**29.** 시계열에 관한 설명 중 틀린 것은?

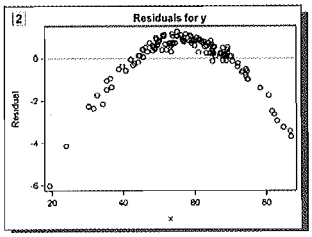
① 대부분의 시계열은 비정상 자료이다. 그러므로 비정상 자료를 정상성 조건에 만족시켜 정상 시계열로 만든 후 시계열 분석을 한다.

② 시계열이 정상 시계열인지 비정상 시계열인지 판단하기 위해 폭발적인 추세를 보이거나 시간에 따라 분산이 변화하는지 관찰해야 한다.

③ 비정상 시계열은 정상 시계열로 변경하고자 할 때 변환과 차분의 방법을 사용한다.

④ 일반적으로 평균이 일정하지 않은 비정상 시계열은 변환을 통해, 분산이 일정하지 않은 비정상 시계열은 차분을 통해 정상 시계열로 바꾼다.

30. 아래의 잔차도를 보고 회귀분석의 가정 중 어떤 가정이 위배되었다고 판단할 수 있는가?




① 비상관성
② 등분산성
③ 선형성
④ 독립성

31. 다음 headsize 데이터는 25개 가구에서 첫 번째와 두 번째 성인 아들의 머리길이(head)와 머리 폭(breadth)을 보여준다. 이에 대한 설명 중 가장 부적절한 것은?

```
> head(headsize)
     head1 breadth1 head2 breadth2
[1,]   191      155   179      145
[2,]   195      149   201      152
[3,]   181      148   185      149
[4,]   183      153   188      149
[5,]   176      144   171      142
[6,]   208      157   192      152
> str(headsize)
 num [1:25, 1:4] 191 195 181 183 176 208 189 197 188 192 ...
 - attr(*, "dimnames")=List of 2
  ..$ : NULL
  ..$ : chr [1:4] "head1" "breadth1" "head2" "breadth2"
> out<-princomp(headsize)
> print(summary(out),loadings=TRUE)
Importance of components:
                          Comp.1  Comp.2   Comp.3  Comp.4
Standard deviation        15.1     5.42     4.12    3.000
Proportion of Variance    0.8      0.10     0.06    0.032
Cumulative Proportion     0.8      0.91     0.97    1.000

Loadings:
           Comp.1   Comp.2   Comp.3   Comp.4
head1      0.570     0.693   -0.442
breadth1   0.406     0.219    0.870   -0.173
head2      0.601    -0.633   -0.209   -0.441
breadth2   0.386    -0.267             0.881
```

① 주성분분석의 결과를 보여준다.
② 첫 두 개의 주성분으로 전체 데이터 분산의 91%를 설명할 수 있다.
③ 두 번째 주성분은 네 개의 원변수와 양의 상관관계를 가진다.
④ 네 개의 주성분을 사용하면 전체 데이터 분산을 모두 설명할 수 있다.

32. 데이터 마이닝의 활용 예가 아닌 것은 어느 것인가?

① 병원에서 환자 데이터를 이용해 해당 환자에게 발생 가능성이 높은 병을 예측한다.
② 웹사이트에 접속한 고객 정보를 활용해 고객에게 맞는 상품과 서비스를 추천한다.
③ 대용량 데이터를 통해 선거의 후보자 인지율 확인을 위한 전화조사에 활용할 대상 리스트를 만들어 낸다.
④ 은행에서 대출 심사를 할 때, 고객 데이터를 활용해 고객의 우량/불량을 예측한다.

33. 데이터 마이닝 모델링 방법 중 분류(Classification) 방법으로 활용되지 않는 R 패키지는 무엇인가?

① rpart ② kmeans ③ party ④ marginTree

34. 모형의 성능을 평가할 때 사용되는 방법론 중 사후확률과 각 분류 기준값에 의해 오분류 행렬을 만든 다음, 민감도(Sensitivity)와 특이도(Specificity)를 산출하여 도표에 도식화하여 평가하는 방식은 무엇인가?

① ROC(Receive Operating Characteristics)
② 이익도표(Lift)
③ AUROC
④ 예측률(Prediction Rate)

35. K-means 군집분석과 계층적 군집분석의 차이를 잘못 설명한 것은?

① K-means 군집분석은 계층적 군집분석과는 달리 한 개체가 처음 속한 군집에서 다른 군집으로 이동해 재배치될 수 있다.
② K-means 군집분석은 초기값에 대한 의존이 커서 초기값을 어떻게 하느냐에 따라 군집이 달라질 수 있다.
③ K-means 군집분석은 동일한 거래계산법을 적용하면 몇 번을 시행해도 동일한 결과가 나온다.
④ 계층적 군집분석은 동일한 거리계산법을 적용하면 몇 번을 시행해도 동일한 결과가 나온다.

36. 데이터를 이용해 분석한 결과 "샌드위치를 사는 고객의 30%가 탄산수를 함께 산다"와 같은 결과를 얻기 위해 실행되는 데이터 마이닝 분석 방법론은 무엇인가?

① 군집분석(Clustering)
② 분류분석(Classification Analysis)
③ 장바구니분석(Market Basket Analysis)
④ 순차분석(Sequence Analysis)

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

① 장바구니분석 시스템

② 부정사용방지 시스템

③ 데이터 마트

④ 교차판매 시스템

**38.** 아래는 스위스의 47개 프랑스어 사용지역의 출산율(Fertility)과 관련된 변수들을 사용하여 얻은 결과이다. 회귀모형에 관한 다음 설명 중 가장 부적절한 것은?

```
> summary(lm(Fertility~., data=swiss))

Call:
lm(formula = Fertility ~ ., data = swiss)

Residuals:
     Min       1Q   Median       3Q      Max
-15.2743  -5.2617   0.5032   4.1198  15.3213

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      66.91518   10.70604   6.250 1.91e-07 ***
Agriculture      -0.17211    0.07030  -2.448  0.01873 *
Examination      -0.25801    0.25388  -1.016  0.31546
Education        -0.87094    0.18303  -4.758 2.43e-05 ***
Catholic          0.10412    0.03526   2.953  0.00519 **
Infant.Mortality  1.07705    0.38172   2.822  0.00734 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.165 on 41 degrees of freedom
Multiple R-squared:  0.7067,    Adjusted R-squared:  0.671
F-statistic: 19.76 on 5 and 41 DF,  p-value: 5.594e-10
```

① 유의수준 0.05하에서 위의 회귀모형은 유의적으로 출산율을 설명한다.

② 위의 설명변수들은 출산율 변동의 원인임을 보여준다.

③ 위의 회귀모형은 출산율 변동의 70.67%를 설명한다.

④ 수정결정계수는 0.671이다.

**39.** 아래 데이터 셋 A, B 간의 유사성을 유클리드 거리로 계산하면?

| 구 분 | A | B |
|---|---|---|
| 키 | 180 | 175 |
| 몸무게 | 65 | 70 |

① 0 ② $\sqrt{5}$ ③ $\sqrt{25}$ ④ $\sqrt{50}$

40. 분류문제를 예측하기 위한 모형을 개발하여 테스트 데이터를 통해 그 결과를 분석하고자 한다. 아래 표를 활용하여 민감도를 구하려고 할 때 민감도를 산출하는 방식은 어떤 것인가?

| | | 질병 여부 | |
|---|---|---|---|
| | | 양성 | 음성 |
| 테스트 | 양성 | TP(True Positive) | FP(False Positive) |
| | 음성 | FN(False Negative) | TN(True Negative) |

① TP/(TP+FN) ② FN/(TP+FN) ③ FP/(FP+TN) ④ TN/(FP+TN)

주관식 - 10문항

01. 개인의 사생활 침해를 방지하고 통계 응답자의 비밀사항은 보호하면서 통계자료의 유용성을 최대한 확보 할 수 있는 데이터변환 방법은 무엇인가?

02. 아래의 ( ㄱ )에 적절한 데이터베이스 용어는?

> 데이터( ㄱ )이란 데이터베이스 내의 데이터에 대한 정확성 일관성, 유효성, 신뢰성을 보장하기 위해 데이터 변경 혹은 수정 시 여러 가지 제한을 두어 데이터의 정확성을 보증하는 것을 말한다.

03. 데이터 분석 기획을 위해서 데이터 분석 수준진단이 필요하다. 분석 준비도와 분석 성숙도를 통해 데이터 분석 수준을 진단하게 되는데, 분석준비도 6개의 영역 중 2가지를 적으시오.

04. 아래 (ㄱ), (ㄴ), (ㄷ) 안에 들어갈 용어를 순서대로 기입하시오.

> 비즈니스 모델 캔버스는 9가지 블록을 단순화하여 ( ㄱ ), ( ㄴ ), 고객단위로 문제를 발굴하고 이를 관리하는 규제와 감사, ( ㄷ ) 영역으로 나눠 분석 기회를 도출한다.

05. 이것은 데이터 안의 두 변수 간의 관계를 알아보기 위해 사용하는 값이다. 두 변수간의 공분산으로는 음과 양의 관계를 파악할 수 있으나 관계 정도를 확인하기는 힘들다. 그래서 각 변수의 공분산을 표준편차의 곱으로 나누어 -1에서 1사이 값으로 표준화하여 두 변수 간의 관계 정도를 확인할 수 있도록 수치화 한 이것을 활용한다. 이것은 무엇인가?

06. 아래 R 코드의 출력 결과는?

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

07. 우리는 모집단을 조사하기 위해 추출한 모집단의 일부 원소를 이용한다. 통계자료의 획득 방법 중 모집단을 조사하기 위해 추출한 집단을 무엇이라 하는가?

08. 다음 중 아래 거래 전표에서 연관 규칙 "C→A"의 신뢰도를 구하시오.

| 물 품 | 거래건수 |
|---|---|
| {A} | 100 |
| {C} | 50 |
| {A, C} | 200 |
| {B, C} | 250 |
| {B, D} | 200 |
| {A, B, D} | 200 |
| {A, B, C, D} | 100 |

09. 동시에 구매될 가능성이 큰 상품군을 찾아내는 연관성 측정에 시간이라는 개념을 포함시켜 순차적인 구매 가능성이 큰 상품군을 찾아내는 데이터 마이닝 기법은?

10. 주성분분석을 통해 얻은 R 프로그램의 결과가 아래와 같이 나왔다. 3개의 변수를 활용할 경우 전체 데이터의 몇 %를 설명할 수 있는지 쓰시오.(소수점 셋째 자리에서 반올림하여 표기하시오.)

```
> summary (princomp ( g, cor=TRUE) )
Timprtance of components :

                        Comp.1     Comp.2     Comp.3     Comp.4     Com.5      Com.6
Standard deviation      1.4836004  1.2189318  0.9559265  0.8630097  0.7066937  0.39364935
Proportion of Variance  0.3668747  0.2476324  0.1522993  0.1241310  0.0832360  0.02582664
Cumulative Proportion   0.3668747  0.6145071  0.7668064  0.8909374  0.9741734  1.00000000
```

# 제 1회 모의고사 답안

【 객관식 정답 】

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 01 | ④ | 11 | ② | 21 | ④ | 31 | ③ |
| 02 | ② | 12 | ④ | 22 | ② | 32 | ③ |
| 03 | ① | 13 | ① | 23 | ① | 33 | ② |
| 04 | ③ | 14 | ③ | 24 | ③ | 34 | ① |
| 05 | ④ | 15 | ② | 25 | ④ | 35 | ③ |
| 06 | ④ | 16 | ① | 26 | ④ | 36 | ③ |
| 07 | ① | 17 | ② | 27 | ② | 37 | ② |
| 08 | ② | 18 | ③ | 28 | ② | 38 | ② |
| 09 | ② | 19 | ① | 29 | ④ | 39 | ④ |
| 10 | ③ | 20 | ④ | 30 | ② | 40 | ① |

【 주관식 정답 】

| | |
|---|---|
| 01 | 마스킹(Masking) |
| 02 | 무결성(Integrity) |
| 03 | 분석업무, 분석 인력/조직, 분석 기법, 분석 데이터, 분석 문화, 분석 인프라 |
| 04 | ㄱ : 업무, ㄴ : 제품, ㄷ : 지원 인프라 |
| 05 | 상관계수(Correlation) |
| 06 | 41 |
| 07 | 샘플(Sample) |
| 08 | 50% |
| 09 | 순차 분석(Sequence Analysis) |
| 10 | 76.68% |



## <과목1. 데이터 이해 - 8문항>

01. 데이터베이스의 일반적인 특징 4가지는 통합된 데이터, 저장된 데이터, 공용 데이터, 운영 데이터이다.

02. 상호 연결된 정보 패턴을 이해하여 이를 토대로 예측한 결과물인 지식을 의미한다.

03. 데이터 사이언티스트에게 요구되는 소프트 역량은 창의적 사고, 호기심, 스토리텔링, 커뮤니케이션 등이 있다.

04. 클라우드 컴퓨팅의 보편화는 빅데이터의 처리 비용을 획기적으로 낮춰 경제성을 제공했다.

05. SQL을 통해 특정 데이터셋에서 필요한 데이터를 추출하고자 할 때 명령문은 "select(특정 변수들) from(데이터 셋) where(조건절 형태의 표현)"으로 사용하면 된다. 조건절 형태의 표현에서 ①은 'A%'로 표현되고, ②는 where 절에서 나올 수 없는 답이며 ③은 '%A%'로 표현하여야 한다.

06. 사생활 침해에 대한 설명이다.

07. 빅데이터 관점에서 사물인터넷은 사물에서 나오는 데이터를 활용해 더욱 지능화된 기기 활용을 할 수 있도록 데이터를 수집하여야 하므로 모든 사물에서 데이터를 추출할 수 있어야 한다.

08. 대량 생산을 통한 가격 경쟁력 확보와 글로벌 네트워크를 통한 판매 확대는 과거의 비즈니스 전략이다.

## <과목2. 데이터 분석 기획 - 8문항>

09. 분석 대상이 명확하게 무엇인지 모르면서 기존 분석 방법으로 새로운 분석을 수행하는 방식은 통찰(Insight)을 도출해 문제의 도출 및 해결에 기여한다.

10. 분석 기획 시 고려사항은 분석의 기본이 되는 데이터에 대한 고려, 활용 가능한 유즈케이스 탐색, 분석 수행에 있어 발생하는 장애요소들에 대한 사전 계획 수립이다.

11. 하향식 접근법은 문제탐색 → 문제정의 → 해결방안 탐색 → 타당성 검토로 전개된다.

12. 시장 니즈 탐색 관점에서 고객 니즈의 변화는 고객, 채널, 영향자들에 의해 진행된다.

13. Political(정치영역)은 주요 정책 방향, 정세, 지정학적 동향 등의 거시적인 흐름을 토대로 분석 기회를 도출한다.

14. 분석 프로젝트 영역별 주요 관리 항목에는 범위, 시간, 원가, 품질, 통합, 조달, 자원, 리스트, 의사소통, 이해관계자 등이 있다.

15. 분석과제 중에 발생된 시사점과 분석 결과물이 풀(Pool)로 관리하고 공유된다. 확정된 분석과제는 풀(Pool)로 관리하지 않는다.

16. 기능구조는 별도 분석조직이 없고 해당 업무부서에서 분석을 수행한다. 전사적 핵심분석이 어려우며, 부서 현황 및 실적 통계 등 과거 실적에 국한된 분석 수행 가능성이 높다.

## <과목3. 데이터 분석 개요 - 24문항>

17. 모형을 개발하기 위해서는 학습 데이터와 테스트 데이터로 구분을 해서 학습 데이터로 모델을 개발하고 테스트 데이터로 모델의 적중률을 확인한다. 학습 데이터를 너무 과대하게 학습한 경우, 과대 적합의 문제가 발생하여 테스트 데이터의 적중률은 떨어지고 일반화하기 힘들어 진다.

18. 구간 척도는 측정 대상이 갖고 있는 속성의 양을 측정하는 것으로 구간이나 구간사이의 간격이 의미가 있는 자료(온도, 지수)이다.

19. R코드를 수행하면 다음과 같은 결과가 콘솔창에 출력된다.

```
> c(2,4,6,8) + c(1,3,5,7,9)
[1]  3  7 11 15 11
경고메시지(들):
In c(2, 4, 6, 8) + c(1, 3, 5, 7, 9) :
두 객체의 길이가 서로 배수관계에 있지 않습니다.
```

20. 표본의 분산은 카이제곱분포를 따른다.

21. 다중회귀분석에서 변수선택법은 전진선택법, 후진제거법, 단계적 선택법이 있다.

22. '이상치'라고 규정한 자료는 분석에서 제외를 할 수 있지만 무조건적으로 제거할 수는 없다.

23. 연속형 확률분포의 종류를 묻는 문제로 이항분포는 이산형 확률분포이고, 정규분포, T분포, F분포는 연속형 확률분포이다.

24. 모집단을 군집으로 구분하고 선정된 군집의 원소를 모두 샘플로 추출하는 다단계 추출 방법은 집락추출법의 설명이다.

25. 자료가 추출된 모집단의 분포에 아무 제약을 가하지 않고 검정을 실시하는 방법이 비모수 검정이며 비모수 검정방법에는 부호검정, 윌콕슨의 순위합검정, 맨-휘트니의 U검정, 런 검정, 스피어만의 순위상관계수 등이 있다.

26. R에서 상관계수를 구하기 위해서는 rcor()함수가 아닌 cor()함수 또는 rcorr()함수를 사용하여야 한다. rcorr()함수를 사용하면 type인자를 통해 피어슨과 스피어만 상관계수를 선택할 수 있다.

27. 종속변수에 미치는 영향이 적더라도 독립변수가 추가되면 결정계수는 변한다.

28. 시계열 자료는 대부분이 비정상 자료이며 이런 경우 비정상 자료를 정상성 조건을 만족시켜 정상 시계열로 만든 후 시계열 분석을 실시한다.

29. 일반적으로 평균이 일정하지 않은 비정상 시계열은 차분을 통해, 분산이 일정하지 않은 비정상 시계열은 변환을 통해 정상 시계열로 바꾼다.

30. 회귀분석의 가정은 선형성, 등분산성, 독립성, 비상관성, 정규성이 있다. 아래의 그림은 회귀모형이 등분산성을 위배했다고 판단할 수 있다.

31. 두 번째 주성분은 head2 변수와 breadth2 변수에 대해 음의 상관관계를 가진다.

32. ③번 보기는 데이터 마이닝이 아닌 표본 추출방법에 대한 내용이다.

33. R에서 지원하는 분류(Classification) 방법으로는 rpart, rpartOrdinal, randomForest, party, tree, marginTree, MapTree등 다양한 방법이 있다.

34. ROC 도표는 구축한 모형의 성능을 사후확률과 각 분류 기준값에 의해 오분류 행렬을 만든 다음, x축은 1- 특이도로 y축은 민감도로 설정하여 그려지는 모형을 평가하는 지표이다.

35. k-평균법은 계층적 군집방법과는 달리 한 개체가 속해 있던 군집에서 다른 군집으로 이동해 재배치가 가능하다. 초기값에 의존하는 방법으로 군집의 초기값 선택에 따라 최종 군집이 변할 수 있다.

36. 장바구니에 함께 구매한 상품 데이터를 이용해 분석한 결과 '아메리카노를 마시는 손님 중 10%가 브라우니를 먹는다', '샌드위치를 먹는 고객의 30%가 탄산수를 함께 마신다'와 같은 결과를 얻어내는 방법론을 연관성 분석(장바구니 분석)이라고 한다.

37. 이상값을 검색하여 한 집단에서 매우 크거나, 매우 작으면 의심되는 대상이므로 부정사용방지 시스템에 활용이 가능하다.

38. 위의 설명변수들은 출산율 변동에 영향을 미치는지 확인하기 위한 변수로서 영향을 미치지 않는 변수도 포함되어 있다. Examination 같은 경우는 출산율 변동의 원인이 아니다.

39. 유클리드 거리를 구하는 공식은 $d(x,y)=\sqrt{(x_1-y_1)^2+\cdots+(x_p-y_p)^2}=\sqrt{(x-y)'(x-y)}$ 이다.
아래의 값을 대입하면 $\sqrt{(180-175)^2+(65-70)^2}=\sqrt{50}$ 이다.

40. 분석결과를 확인하는 방법 중 민감도를 구하는 방법
민감도 : 양성이라고 판단되는 값(TP)/실제 양성의 값(TP+FN)
정확도 : 양성, 음성이라고 판단되는 값(TP+TN)/실제 양성과 음성의 값(TP+FN+FP+TN)
특이도 : 음성이라고 판단되는 값(TN)/실제 음성의 값(FP+TN)

# 데이터 분석 준전문가 모의고사(ADsP) 2회

**출제** 데이터에듀
**문항수** 객관식 : 40 / 주관식 : 10

▶ 배점 안내 : 객관식(40문항) 각 2점 / 주관식(10문항) 각 2점

과목1. 데이터 이해 - 8문항

**01.** 빅데이터 시대에는 데이터를 많이 확보했거나 확보할 수 있는 기업이 혁신을 시도하거나 경쟁력과 생산성 향상을 도모하기에 유리하다. 다음 보기 중 이러한 속성에 부합되기 어려운 기업 분류는?

① 신용카드회사 ② 여행회사 ③ B2B기업 ④ 이동통신사

**02.** 데이터와 정보의 차이를 구분하는 것은 중요하다. 다음 중 정보에 대한 예로 가장 부적절한 것은?

① 평균 구매액 ② 주문 수량 ③ 베스트셀러 ④ 우량 고객

**03.** 영화 '마이너리티 리포트'에 나오는 것처럼 범죄 예측 프로그램에 의해 범행을 저지르기 전에 체포될 수도 있는 사례를 통해 알 수 있는 빅데이터 시대의 위기 요인으로 적절한 것은?

① 소셜 네트워크 ② 사생활 침해 ③ 데이터 오용 ④ 책임 원칙 훼손

**04.** 미국을 의미하는 'The Unites States'는 미국의 남북전쟁이 발발하기 전까지는 아메리카 대륙의 주(州)들이 연합이라는 의미로 복수로 취급되었다는 것을 구글의 'Ngram Viewer'를 통해 확인할 수 있었는데 이와 같이 빅데이터에 거는 기대를 표현한 것은 어느 것인가?

① 산업혁명의 석탄, 철
② 21세기의 원유
③ 렌즈
④ 플랫폼

**05.** 다음 중 빅데이터의 가치 산정이 어려운 이유의 사례로 보기 어려운 것은?

① 전기차 배터리 정보를 충전소 최적지 선정과 같은 2차적 목적에 활용
② 은행 대출심사 알고리즘 작동 원리 이해의 어려움
③ 구글 검색에서 나타나는 것과 같은 데이터의 반복적 재사용
④ 독자의 전자책 독서 순서 정보가 저자의 글쓰기 방식에 영향을 주는 현상

**06.** 전략적 분석을 통해 놀라운 성과를 얻은 미국의 최대 카지노 관련 회사인 하라스엔터테인먼트의 회장인 러브먼은 분석 기반 경영이 도입되지 못하는 이유를 이야기하였다. 보기에서 그 내용이 아닌 것은?

① 기존 관행을 그냥 따를 뿐 중요한 시도를 하지 않는다.
② 경영진이 의사결정 시 직관적으로 결정했을 때 성과가 나오는 것을 경영진의 진정한 재능이라고 생각한다.
③ 분석적 실험을 능숙하게 해내는 사람이 많지 않고 적절한 방법조차 제대로 익히지 못한 사람이 분석업무를 한다.
④ 사람들은 아이디어를 낸 사람이 누군지 보다는 아이디어 자체에 관심을 더 많이 가지고 있다.

**07.** 인문학 열풍 중 최근 사회경제적 환경의 변화로 아닌 것은?

① 복잡한 세계화에서 단순한 세계화로 변화했다.
② 비즈니스의 중심이 제품생산에서 서비스로 이동되었다.
③ 경제와 산업의 논리가 생산에서 시장창조로 바뀌었다.
④ 기존 사고의 틀을 벗어나 문제를 바라보고 창의적으로 문제를 해결하는 능력이 요구되고 있다.

**08.** 다음 개인정보 비식별화 기술 중 아래에서 설명하고 있는 것으로 가장 적절한 것은?

| • 개인정보의 주요 식별요소를 다른 값으로 대체하여 개인 식별을 어렵게 만드는 기술 |
|---|

① 가명처리(Pseudonymization)
② 데이터삭제(Data Reduction)
③ 범주화(Data Suppression)
④ 데이터마스킹(Data Masking)

**09.** KDD 분석 절차 중 분석 목적에 맞는 변수를 찾고 데이터 차원을 축소하는 과정은?

① 데이터셋(Selection)
② 데이터 전처리(Processing)
③ 데이터 변환(Transformation)
④ 데이터 마이닝(Data Mining)

**10.** 데이터 분석 방법론 중 CRISP-DM에 대한 설명으로 옳지 않은 것은?

① 1996년 유럽연합의 ESPRIT에서 있었던 프로젝트에서 시작되어 SPSS, NCR, Daimler Chrysler 등이 참여하였다.
② 각 단계는 폭포수 모델처럼 구성되어져 있다.
③ 모델링 과정에서 데이터셋이 추가로 필요한 경우 데이터 준비 단계를 반복 수행할 수 있다.
④ CRISP-DM은 계층적 프로세스 모델로써 4레벨로 구성되어 있다.



**11.** 비즈니스 모델 캔버스를 활용한 과제 발굴 영역에 대한 설명으로 옳지 않은 것은?

① 업무 : 제품 및 서비스를 생산하기 위해서 운영하는 내부 프로세스 및 주요 자원 관련 도출
② 제품 : 생산 및 제공하는 제품·서비스를 개선하기 위한 관련 주제 도출
③ 고객 : 제품·서비스를 제공받는 사용자 및 고객, 이를 제공하는 채널의 관점에서 관련 주제 도출
④ 규제와 감사 : 분석을 수행하는 시스템 영역 및 이를 운영·관리하는 시스템의 관점에서 주제 도출

**12.** 분석 과제를 도출하기 위한 상향식 접근방식에 대한 설명으로 옳지 않은 것은?

① 상향식 접근방식의 데이터 분석은 비지도 학습방법에 의해 수행된다.
② 분석적으로 사물을 인식하려는 'Why'관점에서 접근한다.
③ 인과관계로부터 상관관계분석으로의 이동이라는 변화를 만들었다.
④ 사물을 있는 그대로 인식하는 'What'관점에서 접근한다.

**13.** 분석과제의 주요 관리 영역이 아닌 것은?

① Data Size
② Data Complexity
③ Speed
④ Analytic & Accessibility

14. 마스터 플랜 수립 시점에서 데이터 분석의 지속적인 적용과 확산을 위한 거버넌스 체계의 구성 요소가 아닌 것은?

① Process ② System ③ Organization ④ Data Resource

15. 기업의 데이터 분석 수준을 진단하는 과정에서 기업에 필요한 6가지 분석 구성요소를 갖추고 있고, 현재 부분적으로 도입되어 지속적인 확산이 필요한 기업들의 분석 수준을 포트폴리오 사분면으로 정의한다면 어디에 해당하는가?

① 준비형 기업 ② 도입형 기업 ③ 정착형 기업 ④ 확산형 기업

16. 다음 중 분석 프로젝트 관리에 대한 설명으로 가장 부적절한 것은?

① 분석 프로젝트 관리는 프로젝트관리 지침(KSA ISO 21500:2013)을 가이드로 활용할 수 있다.
② 데이터 분석 모델의 품질을 평가하기 위해서 SPICE를 활용할 수 있다.
③ 분석 프로젝트의 일정계획 수립 시 데이터 수집에 대한 철저한 통제와 관리가 필요하다.
④ 분석 프로젝트의 최종 결과물이 분석 보고서 형태 또는 시스템인지에 따라 프로젝트 관리에 차이가 있다.

## 과목3. 데이터 분석 - 24문항

17. 다음 중 이산형 확률분포에 해당하지 않는 것은?

① 기하 분포 ② 이항 분포 ③ 지수 분포 ④ 초기하 분포

18. 다음 중 R에서 사용 가능한 데이터 오브젝트에 관한 설명으로 가장 부적절한 것은?

① 차원을 가진 벡터를 행렬이라고 한다.
② 리스트에서 원소들은 다른 모드여도 상관없다.
③ 벡터에서 모든 원소는 같은 모드여야 한다.
④ 데이터프레임은 테이블로 된 데이터 구조로써 행렬로 표현된다.

19. 다음 중 결과가 다른 R코드는?

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

**20.** 종속변수를 설명하는데 가장 중요한 독립변수로 적절한 것은?

① p-value가 가장 작은 변수
② 표준화 자료로 추정한 계수가 가장 큰 변수
③ 원 자료로 추정한 계수가 가장 큰 변수
④ 종속변수와의 상관관계분석에서 상관계수가 가장 큰 변수

**21.** 아래 거래 전표에서 연관성 규칙 A → B 일 때의 지지도는?

| 물 품 | 거래건수 |
|---|---|
| {A} | 10 |
| {B} | 5 |
| {C} | 25 |
| {A, B, C} | 5 |
| {B, C} | 20 |
| {A, B} | 20 |
| {A, C} | 15 |

① 15%　② 20%　③ 25%　④ 30%

29 DAY

**22.** 다음 중 중심극한정리(Central Limit Theorem)에 대한 설명으로 가장 부적절한 것은?

① 여러 통계적 방법론에는 정규데이터가 필요하지만 중심극한정리를 사용하면 비정규적인 모집단에도 이와 유사한 절차를 적용할 수 있다.
② 표본평균의 분포는 표본의 크기가 커짐에 따라 정규분포로 근사한다.
③ 모집단의 분포가 정규분포에 가까워져야 표본평균의 분포가 정규분포로 근사하게 된다.
④ 모집단의 분포가 대칭이면 표본의 크기가 작아도 되지만 모집단의 분포가 비대칭이면 표본의 크기가 30이상이 되어야 한다.

**23.** 다음은 데이터의 척도에 관한 설명이다. 설명이 틀린 것은?

① 명목척도는 측정 대상이 어느 집단에 속하는지 분류할 때 사용되며, 성별, 출생지 정보가 해당된다.
② 순서척도는 측정 대상이 순서를 갖는 자료를 의미하며, 만족도, 선호도, 학년, 신용등급 정보가 해당된다.
③ 구간척도는 측정 대상의 순서와 순서 사이의 간격이 의미가 있는 자료를 의미하며, 온도, 물가지수, 주가지수 정보가 해당된다.
④ 비율척도는 측정대상의 값이 비율로 정의되는 자료를 의미하며, 물가성장율, 흡연감소율의 정보가 해당된다.

**24.** 다음은 확률변수에 관한 설명이다. 설명이 옳지 않은 것은?

① 확률변수는 특정값이 나타날 가능성이 확률적으로 주어지는 변수이며, 실수값으로 표현된다.
② 이산형 확률변수는 확률변수의 공간이 유한하거나 셀 수 있는 경우를 의미하며, 이항분포, 기하분포, 다항분포가 해당된다.
③ 연속형 확률변수는 확률변수의 공간이 무한한 경우를 의미하며, 베르누이 확률분포, 포아송 분포, 정규분포가 해당된다.
④ 균일분포는 확률변수의 구간[a, b] 내에서 모든 확률이 동일한 분포를 의미하며, 확률은 1/(b-a)가 된다.

**25.** 회귀분석에서 변수 선택법에 대한 설명으로 가장 부적절한 것은?

① 전진선택법은 중요하다고 생각되는 설명변수부터 차례로 선택하는 방법이다.
② 전진선택법과 후진선택법의 결과가 항상 동일하지는 않다.
③ 모든 가능한 회귀모형은 독립변수들의 조합으로 이루어진 회귀모형 중 가장 적합하게 나타난 모형을 선택하는 방법이다.
④ 전진선택법으로 변수를 추가할 때 기존 변수들의 중요도는 영향을 받지 않는다.

**26.** 분해시계열에 대한 설명 중 잘못된 것은?

① 분해시계열이란 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법을 말한다.
② 분해 시계열의 분해 요소는 추세요인, 계절요인, 순환요인, 회귀요인으로 크게 4가지로 이루어진다.
③ 추세요인은 자료의 형태가 오르거나 내리는 추세를 따르는 경우로 선형적 형태, 지수형태 등이 있다.
④ 순환요인은 경제적이나 자연적인 이유가 없이 알려지지 않은 주기를 가지고 변화하는 자료 형태이다.

**27.** 두 개 이상의 독립변수를 사용해 하나의 종속변수의 변화를 설명하는 다중회귀분석을 실시할 것이다. 다음 중 모형을 적합 시킨 후, 모형이 적절한지 확인하기 위해 체크해야 할 사항으로 부적절한 것은?

① 상관계수를 통해 모형의 설명력을 확인한다.
② F-value를 통해 모형이 통계적으로 유의한지 확인한다.
③ 모형이 데이터에 잘 적합되어 있는지를 확인한다.
④ t-value, p-value를 통해 유의한지 확인한다.

**28.** 주성분분석은 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 것이 목적이다. 다음 중 주성분분석에 대한 설명으로 적절하지 않은 것은 무엇인가?

① 다변량 자료를 저차원의 그래프로 표시하여 이상치(Outlier) 탐색에 사용한다.

② 변수들끼리 상관성이 있는 경우, 해석상의 복잡한 구조적 문제가 발생하는데 이를 해결하기 위해 사용한다.

③ 회귀분석에서 다중공선성(Multicollinearity)의 문제를 해결하기 위해 활용한다.

④ p개의 변수들을 중요한 m(p)개의 주성분으로 표현하여 전체 변동을 설명하는 것으로 m개의 주성분은 원래 변수와는 관계없이 생성된 변수들이다.

**29.** 아래는 데이터프레임 mtcars를 이용해 회귀분석을 수행한 R 명령의 결과이다. 다음 중 이 결과에 대한 설명으로 가장 부적절한 것은?

```
> summary(lm(mpg~., data=mtcars))

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633 .
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

① 오차의 표준편차 추정치는 2.65이다.

② 모든 독립변수가 유의수준 0.1에서 유의하지 않다.

③ 후진제거법을 적용할 때 가장 먼저 제거될 독립변수는 cyl이다.
④ 유의수준 0.01하에서 이 회귀모형은 유의하다.

**30.** 데이터 마이닝을 위한 데이터 분할에 대한 설명으로 틀린 것은 어느 것인가?

① 데이터를 구축용(Training), 검정용(Validation), 시험용(Test)으로 분리한다.
② 일반적으로 데이터 구축용, 검정용, 시험용 데이터는 50%, 30%, 20%로 정한다.
③ 데이터가 충분하지 않을 때는 구축용과 시험용 데이터만 구분하여 활용한다.
④ 통계학에 적용되는 교차확인(Cross-Validation)은 데이터 마이닝에서 활용할 수 없다.

**31.** Default 데이터는 10,000명의 신용카드 고객에 대한 체납 여부(default)와 학생여부(student), 카드 잔고(balance), 연봉(income)을 포함하고 있다. 고객의 체납 확률을 예측하기 위한 아래 결과에 대한 설명으로 가장 부적절한 것은?

```
> summary(glm(default~.,data=Default,family="binomial"))

Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

① 로지스틱 회귀모형을 사용한 결과이다.
② 카드 잔고와 연봉이 동일한 수준일 때, 학생(studentYes)이 학생이 아닌 고객보다 체납확률이 낮다.
③ 세 설명변수 모두 체납확률을 예측하는데 유익한 영향이 있다.
④ 동일한 신분과 연봉 수준일 때 카드 잔고가 높을수록 체납 확률이 높다.

32. 데이터 마이닝 분석 기법 중 의사결정나무 분석의 특성으로 잘못 표현한 것은 어느 것인가?

① 의사결정나무 모형의 결과는 누구나 이해가 쉽고 설명이 용이하다.
② 의사결정나무 알고리즘의 모형 정확도는 다른 분류모형에 뒤지지 않는다.
③ 의사결정나무 알고리즘은 대용량 데이터에서도 빠르게 만들 수 있고 데이터의 분류 작업도 신속히 진행할 수 있다.
④ 의사결정나무 알고리즘은 비정상적인 잡음 데이터에서는 민감하여 분류가 쉽지 않다.

33. 다음 중 비모수적 방법에 대한 설명으로 가장 부적절한 것은?

① 관측된 자료가 주어진 분포를 따른다는 가정을 받아들일 수 없을 때 이용하는 검정법이다.
② 자료가 추출된 모집단의 분포에 대해 제약을 가하지 않고 검정을 실시하는 방법이다.
③ 관측된 자료로 구한 표본평균과 표본분산 등을 이용해 검정을 실시한다.
④ 관측된 자료가 특정 분포를 따른다고 가정할 수 없을 때 이용한다.

34. 비계층적 군집분석의 장점에 대한 설명이 잘못된 것은?

① 주어진 데이터의 내부 구조에 대한 사전 정보가 없어도 의미 있는 결과를 얻을 수 있다.
② 다양한 형태의 데이터의 적용이 가능하다.
③ 분석방법의 적용이 용이하다.
④ 사전에 주어진 목적이 없으므로 결과 해석이 쉽다.

35. 아래의 데이터 마이닝 분석 예제 중 비지도(Unsupervised) 분석을 수행해야 하는 예제는?

| 가. 우편물에 인쇄된 우편번호 판별 분석을 통해 우편물을 자동으로 분류<br>나. 고객의 과거 거래 구매 패턴을 분석하여 고객이 구매하지 않은 상품을 추천<br>다. 동일 차종의 수리 보고서 데이터를 분석하여 차량 수리에 소요되는 시간을 예측<br>라. 상품을 구매할 때 그와 유사한 상품을 구매한 고객들의 구매 데이터를 분석하여 쿠폰을 발행 |
|---|

① 나, 다 ② 가, 라 ③ 가, 다 ④ 나, 라

36. 다음 중 연관분석에서 '항목 A를 포함한 거래 중에서 항목 A와 항목 B가 같이 포함될 확률은 어느 정도인가를 나타내 주는 연관성의 정도'로 정의되는 측도로 가장 적절한 것은?

① 지지도 ② 신뢰도 ③ 특이도 ④ 민감도

37. 데이터 프레임 attitude 아래와 같이 R명령을 적용하고 결과를 얻었다. 다음 설명 중 가장 부적절한 것은?

```
> cor(attitude)
             rating complaints privileges  learning    raises  critical   advance
rating     1.0000000  0.8254176  0.4261169 0.6236782 0.5901390 0.1564392 0.1550863
complaints 0.8254176  1.0000000  0.5582882 0.5967358 0.6691975 0.1877143 0.2245796
privileges 0.4261169  0.5582882  1.0000000 0.4933310 0.4454779 0.1472331 0.3432934
learning   0.6236782  0.5967358  0.4933310 1.0000000 0.6403144 0.1159652 0.5316198
raises     0.5901390  0.6691975  0.4454779 0.6403144 1.0000000 0.3768830 0.5741862
critical   0.1564392  0.1877143  0.1472331 0.1159652 0.3768830 1.0000000 0.2833432
advance    0.1550863  0.2245796  0.3432934 0.5316198 0.5741862 0.2833432 1.0000000
```

① 모든 변수들 사이에 양(+)의 상관관계가 존재한다.

② rating과 complaints 사이에 가장 강한 상관관계가 존재한다.

③ critical과 learning 사이의 상관관계가 가장 약하다.

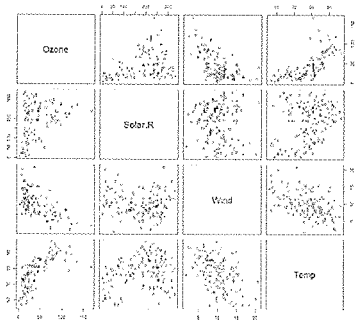
④ 모든 변수의 분산이 1이다.

38. 아래의 데이터는 두 종류의 수면유도제(group)에 대해 무작위로 선정된 20명의 환자를 대상으로 수면 시간의 증감(extra)을 측정한 자료이다. 다음 중 결과에 대한 설명으로 가장 부적절 한 것은?

```
> sleep
    extra group
1    0.7     1
2   -1.6     1
3   -0.2     1
4   -1.2     1
5   -0.1     1
6    3.4     1
7    3.7     1
8    0.8     1
9    0.0     1
10   2.0     1
11   1.9     2
12   0.8     2
13   1.1     2
14   0.1     2
15  -0.1     2
16   4.4     2
17   5.5     2
18   1.6     2
19   4.6     2
20   3.4     2
> summary(sleep$extra)
Min.   1st Qu.   Median    Mean   3rd Qu.    Max.
-1.600  -0.025   0.950     1.540   3.400     5.500
```

① 평균적으로 1.54 시간의 수면시간 증가를 가져왔다.
② 3.4 시간 이상 수면이 증가한 환자는 약 25%이다.
③ 모든 환자들의 수면시간이 증가하였다.
④ 가장 많이 증가한 수면시간은 5.5시간이다

**39.** 아래의 산점도 행렬에 대한 설명으로 가장 부적절한 것은?(변수 : Ozone, Solar.R, wind, temp)




① temp와 wind 간의 관계는 상대적으로 선형이다.
② Solar.R와 ozone의 관계는 명확하지 않다.
③ ozone과 wind 간에는 양의 상관관계가 있다.
④ wind와 Solar.R 간에는 비선형 관계가 있다.

**40.** 다음 중 과대적합(Overfitting)에 대한 설명으로 가장 부적절한 것은?

① 과대적합이 발생할 것으로 예상되면 학습을 종료하고 업데이트하는 과정을 반복해 과대적합을 방지할 수 있다.
② 과대적합은 분석 변수가 너무 많이 존재하고 분석 모델이 복잡할 때 발생한다.
③ 분석 데이터가 모집단의 특성을 설명하지 못하면 발생한다.
④ 생성된 모델은 분석 데이터에 최적화되었기 때문에 훈련 데이터의 작은 변화에 민감하게 반응하는 경우는 발생하지 않는다.

29 DAY

## 주관식 - 10문항

**01.** 인터넷상의 서버에서 데이터 저장, 처리, 네트워크, 콘텐츠 사용 등 서로 다른 물리적인 위치에 존재하는 컴퓨팅 자원을 가상화 기술을 통해 IT 관련 서비스를 한 번에 제공하는 혁신적인 컴퓨팅 기술은 무엇인가?

**02.** 데이터 사이언스란 데이터로부터 의미 있는 정보를 추출하는 학문이다. 통계학이 정형화된 실험 데이터를 분석 대상으로 하는 것에 비해, 데이터 사이언스는 정형 또는(　　)을 막론하고 인터넷, 휴대전화, 감시용 카메라 등에서 생성되는 숫자와 문자, 영상 정보 등 다양한 유형의 데이터를 대상으로 한다.

03. 풀어야 할 문제에 대한 상세한 설명 및 해당 문제를 해결했을 때 발생하는 효과를 명시함으로써 향후 데이터 분석 문제로의 전환 및 적합성 평가에 활용하도록 하는 것은 무엇인가?

04. 아래의 R 명령의 결과를 쓰시오.

```
0/0
```

05. 100명의 키를 cm으로 측정한 데이터의 분산이 225였다. 동일한 100명의 키를 m로 측정한다면 데이터의 분산은 얼마인가?

06. 고객은 늘 구매하지 않는다. 경쟁사의 고객 빼앗기에 따른 고객의 변심 또는 고객의 니즈나 취향이 변해 더 이상 상품과 서비스를 사용하지 않고 경쟁사와 거래하는 고객을 무엇이라고 하는가?

07. 의사결정나무 중 연속형 타깃변수(또는 목표변수)를 예측하는 의사결정나무를 무엇이라고 하는가?

08. 데이터 마이닝 모델링 분석 기법 중 random input에 따른 forest of tree를 이용한 분류방법으로 랜덤한 forest에는 많은 트리들이 생성된다. 새로운 오브젝트를 분류하기 위해 forest에 있는 트리에 각각 투입해 각각의 트리들이 voting함으로써 분류하는 방식의 R 패키지는 무엇인가?

09. 다수 모델의 예측을 관리하고 조합하는 기술을 메타 학습(Meta Learning)이라 한다. 여러 분류기(Classifier)들의 예측을 조합함으로써 분류 정확성을 향상시키는 기법은?

10. 데이터셋 x는 두 개의 변수와 5개의 관측치를 가지며 아래는 데이터와 관측치 간의 유클리드 거리를 나타낸다. 최단연결법을 사용하여 계층적 군집화를 할 때 첫 단계에서 형성되는 군집과 관측치 a와의 거리를 구하시오.

```
> x
  x1 x2
a  1  4
b  2  1
c  4  6
d  4  3
e  5  1
> dist(x)
     a    b    c    d
B  3.2
c  3.6  5.4
d  3.2  2.8  3.0
e  5.0  3.0  5.1  2.2
```

# 제 2회 모의고사 답안

【 객관식 정답 】

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 01 | ③ | 11 | ④ | 21 | ③ | 31 | ③ |
| 02 | ② | 12 | ② | 22 | ③ | 32 | ④ |
| 03 | ④ | 13 | ④ | 23 | ④ | 33 | ③ |
| 04 | ③ | 14 | ④ | 24 | ③ | 34 | ④ |
| 05 | ② | 15 | ④ | 25 | ④ | 35 | ④ |
| 06 | ④ | 16 | ③ | 26 | ② | 36 | ② |
| 07 | ① | 17 | ③ | 27 | ① | 37 | ④ |
| 08 | ① | 18 | ④ | 28 | ④ | 38 | ③ |
| 09 | ③ | 19 | ① | 29 | ② | 39 | ③ |
| 10 | ② | 20 | ② | 30 | ④ | 40 | ④ |

【 주관식 정답 】

| | |
|---|---|
| 01 | 클라우드 컴퓨팅 |
| 02 | 비정형 |
| 03 | 분석 유즈 케이스 |
| 04 | NaN(Not a Number) |
| 05 | 0.0225 |
| 06 | 이탈고객 |
| 07 | 회귀나무(Regression Tree) |
| 08 | 랜덤 포레스트(Random Forest) |
| 09 | 앙상블 기법 |
| 10 | 3.2 |



## <과목1. 데이터 이해 - 8문항>

01. B2B기업은 기업 간의 전자 상거래를 진행하여 지속적인 데이터가 생성이 되기 어렵다. 반면에 B2C기업은 고객을 상대로 하기 때문에 고객의 데이터가 지속적으로 생성이 된다.

02. DIKW 피라미드에서 우량고객, 베스트셀러, 평균 구매액은 정보(Information)에 해당되고 주문수량은 데이터(Data)에 해당된다.

03. 민주주의 국가에서 채택한 형사 처벌은 잠재적 위협이 아닌 명확하게 행동한 결과에 대해 책임을 묻고 있다. 특정인이 빅데이터 분석 결과에 따라 특정한 행위를 할 가능성이 높다는 이유만으로 처벌 받는 것은 행위 결과에 대해서만 책임을 묻는다는 사회 원칙을 크게 훼손할 수 있다.

04. 구글의 'Ngram Viewer'를 통해 우리가 확인하기 힘들었던 부분을 찾을 수 있도록 해주는 빅데이터의 비유는 "렌즈"이다.

05. 빅데이터의 가치 산정이 어려운 이유는 다음과 같다.
    1. 데이터 활용방식 : 재사용, 재조합, 다목적용 개발
    2. 새로운 가치 창출
    3. 분석 기술 발전
    ②은 위의 3가지에 해당하지 않는다.

06. 하라스엔터테인먼트의 회장인 러브먼이 언급한 분석 기반 경영이 도입되지 못하는 이유로 "사람들은 아이디어 자체보다는 아이디어를 낸 사람이 누군지에 더 많이 관심을 가지고 있다."고 이야기했다.

07. 최근의 사회경제적 환경은 단순 세계화에서 복잡한 세계화로 변화하고 있다.

08. 가명처리는 개인정보 주체의 이름을 다른 이름으로 변경하는 기술이다.

## <과목2. 데이터 분석 기획 - 8문항>

09. 데이터 전처리 프로세스를 통하여 분석용 데이터 셋이 편성되면 분석 목적에 맞는 변수를 선택하거나 데이터의 차원을 축소하여 데이터 마이닝을 효율적으로 적용될 수 있도록 데이터셋을 변경하는 프로세스를 데이터 변환이라고 한다.

10. CRISP-DM 프로세스의 각 단계는 폭포수 모델처럼 일방향으로 구성되어 있지 않고 단계 간 피드백을 통하여 단계별 완성도를 높인다.

11. 규제와 감사는 제품 생산 및 전달 과정 프로세스 중에서 발생하는 규제 및 보안의 관점에서 주제를 도출한다.

12. 분석적으로 사물을 인식하려는 'Why' 관점은 일반적으로 사용되고 있는 문제 해결방식인 하향식 접근방식을 말한다.

13. 분석 과제의 주요 관리 영역에는 Data Size, Data Complexity, Speed, Analytic&Complexity, Accuracy&Precision가 있다.

14. 분석 거버넌스 체계 구성요소는 Process(과제 기획/운영 프로세스), System(IT 시스템/프로그램), Organization(분석 기획/관리 및 추진 조직), Data(데이터 거버넌스), Human Resource(분석 관련 교육/마인드 육성 체계)가 있다.

15. 포트폴리오 사분면에서 분석이 현재 부분적으로 도입되어 지속적인 확산이 필요한 기업들은 확산형 기업이라고 정의한다.

16. 분석 프로젝트 관리에서 일정계획 수립 시 데이터 수집에 대한 철저한 통제와 관리보다 분석 범위가 빈번하게 변경되므로 시간이 소요될 수도 있다. 따라서 Time Boxing 기법과 같은 방법으로 일정관리를 진행하는 것이 필요하다.

## <과목3. 데이터 분석 - 24문항>

17. 기하, 이항, 초기하 분포는 이산형 확률분포이다.

18. R에서 사용 가능한 데이터 오브젝트(행렬, 벡터, 데이터프레임, 리스트)에 관한 설명으로 데이터프레임은 테이블로 된 구조인 것은 맞지만 행렬이 아닌 리스트 구조로 구현된다.

19. a<-c(1,10)은 벡터 값으로 1, 10이 나타나지만, 나머지는 1부터 10까지의 수를 보여준다.

20. 다중선형회귀분석의 종속변수를 설명하는 가장 중요한 독립변수는 추정한 계수가 클수록 종속변수에 가장 영향을 많이 미치게 된다. 특히 β0가 없는 표준화된 추정식을 만들게 되면 각 계수의 크기를 더욱 정확히 알 수 있게 된다.

21. 지지도를 구하는 공식은 P(A∩B)이므로 25%가 정답이다.

22. 동일한 확률분포를 가진 독립 확률 변수의 분포는 n이 적당히 크다면(n은 30이상) 정규분포에 가까워진다는 정리이다.

23. 비율척도는 측정대상의 간격에 대한 바율이 의미를 가지는 자료를 의미하고 무게, 나이, 시간, 거리 정보가 해당된다.

24. 베르누이 확률분포, 포아송분포는 이산형 확률분포이다.

25. 다중회귀분석에서 변수 선택법 중 전진선택법은 변수가 추가되면 기존 변수들의 중요도에 영향을 받게 된다. 다시 말해, 변수를 추가했는데 이미 선택된 변수의 유의수준이 높아지면 추가한 변수를 활용하지 못하게 된다.

26. 분해 시계열의 분해 요소는 추세요인, 계절요인, 순환요인, 불규칙요인으로 크게 4가지로 이루어진다.

27. 다중회귀분석의 결과에서 모형의 적절함을 확인하기 위해서는 F 검정 통계량과 유의확률, t 통계량과 유의확률, $R^2$값을 검정해야 한다. 보기 ①과 같이 상관계수를 통해 모형의 설명력을 확인하는 것은 회귀분석 이전의 단계에서 실행해야 한다.

28. p개의 변수들을 중요한 m(p)개의 주성분으로 표현하여 전체 변동을 설명하는 것으로 m개의 주성분은 원래 변수에서 선형결합으로 생성된 변수이다.

29. 다중선형회귀분석 결과, 입력변수 중 wt는 유의수준 0.1하에서 유의하지만 나머지 변수는 유의하지 않다.

30. 필요에 따라서는 구축용과 시험용을 번갈아가며 사용하는 교차확인을 통해 모형을 평가하기도 한다.

31. income은 체납확률을 예측하는데 유의한 변수가 아니다.

32. 의사결정나무 알고리즘은 비정상적인 잡음 데이터에 대해서도 민감함이 없이 분류할 수 있다.

33. 비모수 검정 방법은 모집단의 분포에 대한 아무 제약을 가하지 않고 검정을 실시하고 관측된 자료의 수가 많지 않거나 자료가 개체간의 서열관계를 나타내는 경우 이용한다. 또, 관측된 자료가 특정분포를 따른다고 가정할 수 없는 경우 이용한다. 관측된 자료로 구한 표본평균과 표본분산 등을 이용해 검정을 실시하는 것은 모수적검정 방법이다.

34. 사전에 주어진 목적이 없으므로 결과 해석이 어렵다.

35. 가, 다는 이미 분류된 데이터에 대해 분석을 하므로 지도(Supervised) 분석이다.

36. 지지도는 전체 거래 중 품목 A와 품목 B를 동시에 포함하는 거래의 비율이며, 향상도는 A가 주어지지 않았을 때의 품목 B의 확률에 비해 A가 주어졌을 때의 품목 B의 확률의 증가비율이다.

37. cor(attitude)를 통해서는 분산을 확인하는 것이 아니라 상관계수 값을 확인할 수 있다.

38. 모든 환자들의 수면시간이 증가했다는 결과를 summary(sleep$extra)를 통해서 확인이 불가능하다. mean, 3rd Qu.와 Max의 값으로 나머지는 확인이 가능하다.

39. ozone과 wind 간에는 음의 상관관계가 있다.

40. 생성된 모델이 훈련 데이터에 최적화되어 있기 때문에 테스트 데이터의 작은 변화에 민감하게 반응한다.