R로 배우는 데이터분석 입문 2022년 2학기 중간고사

[유의사항]

- 시험시간: 19:10~20:30(80분)
- 정답이 정수(integer)로 딱 떨어지지 않을 경우 유효 숫자는 소수 셋째자리(소수 넷째자리에서 반올림)로 할 것. 유효숫자 틀리면 부분 감점 있음
- 별도 답안지가 없으니 문제지에 코드는 쓰지 말고, 답만 써서 스노보드에 지정된 시각 안에 제출할 것
- 시험 종료에 임박하면 답안 탑재에 부하가 걸릴 수 있으니 적어도 종료 5분 전에 탑재하도록 하고, 스노보드 탑재 시각 기준으로 늦게 제출하면 감점
- 답안파일명은 안내된 형식(예: 경영학부_210087_오유정)을 반드시 지킬 것
- 시험 도중 zoom 채팅방을 주기적으로 참고하고, 카메라 세팅도 확인할 것
- 불필요한 질문은 삼가고, 필요한 질문만 담당교수에게 직접 채팅으로 할 것(카카오단톡방, 마이크 육성, 전체 채팅 금지)
- 기타 사항은 스노보드 공지사항에 기 공지된 내용을 참조할 것

※ read_csv 함수를 이용하여 churn.csv 파일(이동통신사 고객 관련 데이터)을 불러온 후 (read_csv 함수에서 col_types와 na 조건 입력하지 않아도 됨), 아래 질문에 답하시오.

1. 다음 중 척도가 다른 변수는 무엇인가? (1)
① download ② phone ③ city ④ marriage ⑤ streaming

2. 고객이 가장 많이 거주하는 도시(city)는 어디인가? (Los Angeles)

3. phone, internet, security 세 가지 서비스를 모두 가입한 고객은 몇 명인가?
1736명

4. mean 함수를 이용하여 noreferral의 평균을 구하시오(유효숫자 소수 셋째자리).
1.953

5. 다음 중 측정값의 분포가 정규분포에 가장 가까운 변수는 무엇인가? (2)
① extra data charge ② long charge ③ total contact ④ noreferral ⑤ total service issue

6. population의 중위수(median)는 얼마인가?
17554

7. monthly charge에 대해 0부터 130까지 5단위로 구간을 구분하여 히스토그램을 그렸을 때, 가장 빈도가 많은 구간의 하한값은 얼마인가? 20

8. promotion 측정값이 None이 아닌 고객 중에서 churn category가 Price 혹은 Competitor인 고객은 몇 명인가? 473명

9. 기혼 남성 고객 중에서 cltv가 4,000달러 이상이고 5,000달러 이하인 고객들의 duration 평균을 구하시오(유효 숫자 소수 셋째자리).

26.752

10. monthly charge의 결측치를, 결측치가 제외된 monthly charge의 평균으로 대체한 후, monthly charge가 65달러 이상인 빈도수를 구하시오. 3914

11. noreferral의 결측치를 0으로 변경한 후, noreferral의 값이 0인 고객수를 구하시오.

3821

12. 아래 표를 참고하여 새로운 변수 grade를 만든 후, grade의 네 개 집단 중에서 total service issue 평균이 가장 높은 집단과 해당 집단의 평균(유효 숫자 소수 셋째자리)을 각각 구하시오. C 0.367

조건	grade 측정값
cltv ≤ 3,000	C
3,000 < cltv ≤ 4,000	B
4,000 < cltv ≤ 5,000	A
5,000 < cltv	S

13. churn reason이 Competitor와 관련된 네 가지 중 하나에 해당되는 고객 중에서 download가 60보다 큰 고객은 몇 명인가? 71명

14. city 측정값이 'Santa'를 포함한 고객 중에서 total contact가 가장 많은 고객의 id는 무엇인가? 5520-FVEWJ

15. city 측정값이 'Santa'를 포함한 고객 중에서 extra data charge가 상위 5%에 속한 고객들의 grade에 대해서 아래 괄호안에 빈도수를 채우시오.

S: (7), A: (3), B: (0), C: (0)

16. 고객들이 살고 있는 도시의 종류는 몇 개인가? 1106

17. churn add.csv 파일을 불러와 churn 데이터프레임과 통합한 후, 질문에 답하시오.
(1) unlimited와 payment 두 가지 변수를 동시에 고려하여 고객을 구분했을 때, 빈도수가 가장 적은 집단의 유형과 빈도수는 각각 얼마인가?
유형 : unlimited=Yes payment = 2, 빈도수 : 857
(2) unlimited와 payment 두 가지 변수를 동시에 고려하여 고객을 구분했을 때, population 평균이 가장 큰 집단의 유형과 population 평균은 각각 얼마인가?
유형 : unlimited = Yes payment = 2, 평균 : 22908

18. 문제17에서 통합된 churn 데이터프레임에 있는 계량형 척도로 측정된 변수 중에서 변동계수가 가장 큰 변수와 이때의 변동계수(유효숫자 소수 셋째자리)는 각각 얼마인가? extra data charge, 2.46

19. 문제18에서 구한 변동계수가 가장 큰 변수의 측정값이 '표본평균 ± 3 × 표본표준편차'를 벗어나면 이상치로 판정하려고 한다. 이상치는 모두 몇 개인가?
172개

20. churn 데이터프레임에서 문제19에서 확인된 이상치를 보유한 사례를 모두 제거한 후, promotion에서 'Offer D'를 경험한 고객의 age 평균(유효숫자 소수 셋째자리)을 구하시오. 45.158