[과제 3] Data Manipulation 학습 (20문제)

* 준비사항

첨부된 타이타닉 데이터(titanic.csv)를 불러들여 아래 과제를 수행하시오.

* 기타

- "행을 앞에서부터 n개만 출력"하는 경우, 별다른 지시사항(정렬 등)이 없을 때는 head(n) 함수를 이용하여 그대로 출력하세요.

- dplyr::summarise 함수를 이용하는 경우 결과값의 변수명은 임의로 설정하되, 문제에 명시된 경우에는 명시된 대로 설정하여 출력하세요.

- 평균값을 계산하는 경우, 반드시 NA값을 제거한 후 계산하세요.

1. 데이터의 Name, Sex, Age, Survived 변수만 선택한 후, 앞 5개의 행만 출력하시오.

2. 데이터의 Cabin의 유일한 값들만 가져온 후 앞 5개의 행만 출력하되, 값이 없는 행("")은 제외하시오.

3. 생존 여부(Survived) 별로, 요금(Fare) 평균 값을 변수명 "Fare_mean"으로 설정하여 출력하시오.

4. 살아남은 20세 이하(<= 20)의 승객에 대해 Pclass 별 행 개수와 요금(Fare) 평균을 각각 변수명 "Class_count", "Fare_mean"으로 설정하여 출력하시오. (Pclass가 내림차순이 되도록 정렬하여 출력)

5. 데이터에서 이름에 "th"가 포함(대소문자 상관없이)되는 승객의 수를 구하시오. 또한 해당하는 승객들에 대해 승객번호(PassengerId)와 이름(Name) 변수만 선택하고 앞 5개의 행을 출력하시오.

6. 데이터를 Pclass, Sex로 그룹을 나눈 뒤 각 그룹별 나이(Age)의 평균을 변수명 "Age_mean"으로 설정하여 출력하시오. 이 때, 나이가 내림차순이 되도록 출력하시오.

7. 15세 이하(<= 15)의 승객에 대해 Pclass별로 생존율을 구하려고 할 때, 최종 결과물이 아래와 같은 조건을 만족하도록 출력하시오.

(1) 생존율: 그룹 별 생존(Survived = 1) 승객 수 / 그룹 별 전체 승객 수

(2) 반드시 생존율을 소수점 2의 자리에서 반올림을 한 n.xx% 형식으로 나타낼 것

(3) Pclass가 오름차순이 되도록 정렬할 것

(4) 결과물 예시:

Pclass	Surv_percentage
1	a.xx%
2	b.xx%
3	c.xx%

8. 고객의 정보 및 생존 여부를 하나의 컬럼으로 정리하려고 한다. 먼저 NA값이 포함된 행은 모두 제외하고, Name, Age, Sex, Survived 변수만 사용하여 아래와 같은 변수("Summary")를 만들어 앞 5개 행만 출력하시오.

Summary

[Name: Braund, Mr. Owen Harris | Age: 22 Sex: male | Survived: No]

…(생략)

9. 고객의 특성을 요약한 새로운 변수를 만들고자 한다. 먼저 NA값이 포함된 행은 모두 제외하고, 아래와 같은 조건(좌)에 따라 값(우)을 갖는 새로운 변수"Passenger_char"를 생성하시오.

1) 30세 이상(〉=)인 남성 → 'male_over_30'

2) 30세 미만(〈)인 남성 → 'male_under_30'

3) 30세 이상(〉=)인 여성 → 'female_over_30'

4) 30세 미만(〈)인 여성 → 'female_under_30'

최종적으로 Age, Sex, Passenger_char의 세 변수만 선택하여 앞 5개 행을 출력하시오.

10. 9번에서 만든 변수(Passenger_char)를 이용해, Passenger_char, Pclass별로 사망율을 구하려고 한다. 이 때 최종 결과물이 아래와 같은 조건을 만족하도록 출력하시오.

(1) 사망율: 그룹 별 사망(Survived = 0) 승객 수 / 그룹 별 전체 승객 수

(2) 반드시 사망율을 소수점 2의 자리에서 반올림을 한 n.xx% 형식으로 나타낼 것

(3) 사망율이 내림차순이 되도록 정렬할 것

(4) 결과물 예시:

Pclass	Passenger_char	Death_percentage
?	male_xxx	a.xx%
?	male_xxx	b.xx%

* 준비사항

지정한 링크로 접속하여 2가지 데이터(winequality-red.csv, winequality-white.csv) 를 불러서 아래 과제를 수행하시오.

데이터셋: wine quality dataset (https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/)

11. red wine quality dataset 과 white wine quality dataset 을 불러오고, 각각 type 을 의미하는 column 을 추가한 후, 2개 데이터셋을 하나의 데이터프레임으로 합치시오.

12. 이름 중 'dioxide' 가 포함된 것을 출력할 수 있는 방법 2가지를 r로 작성하시오.

13. density 에 해당하는 값이 density의 평균보다 크고, volatile.acidity가 최대인 조건을 만족하는 데이터를 출력하시오.

14. 전체 데이터에 dioxide_ratio를 새로운 변수로 추가하시오.

참고) dioxide_ratio = free.sulfur.dioxide / total.sulfur.dioxide

(단, 변수명은 dioxide_ratio 로 할 것)

15. quality 가 5인 데이터 중 fixed.acidity의 최대, 최소값에 해당하는 volatile acidity 값을 찾고, 그 결과를 fixed.acidity 의 최대, 최소값과 함께 내림차순으로 정렬한 결과를 출력하시오.

16. 품질기준에 따라 데이터를 분류하고 각 품질별 residual.sugar 의 평균값을 비교하시오.

품질기준:

(1) quality < 6 : Low Quality

(2) 6 <= quality <= 7 : Medium Quality

(3) quality > 7 : High Quality

17.이전 문제에서 제시한 품질 기준에 따라, quality 의 품질 카테고리에 해당하는 **qlabel** column 을 전체 데이터셋에 대해 생성하시오.

품질기준:

(1) quality 〈 6 : Low Quality ➔ qlabel='L'

(2) 6 〈= quality 〈= 7 : Medium Quality ➔ qlabel='M'

(3) quality 〉 7 : High Quality ➔ qlabel='H'

18. 생성한 quality 품질 카테고리(qlabel) 별로 pH 평균과 citric.acid 평균을 출력하시오.

19. 각 와인종류(red, white)에서 각각의 품질 카테고리가 차지하는 비중(ratio) 를 계산하고, 그 결과를 새로운 dataframe으로 출력하시오.

20. volatile.acidity 와 citric.acid의 평균값으로 acid_mean을 구한 후 새로운 변수로 추가하고, dioxide_ratio 가 acid_mean 보다 큰 데이터들을 상위 5개만 보이게 출력하시오.