


데이터 전처리



실습 공지 사항

- 이번 차시 실습부터는, 구름과 colab 을 함께 사용합니다.
 - 실습 **자동 채점**은 구름 에서만 진행됩니다.
(colab 파일은 제출하지 않습니다)
 - 실습 내용에 따라 colab을 이용해 시각화하는 과정이 있습니다.
 - 각 실습 내용을 잘 숙지하여 수행하시기 바랍니다.
- 



1

이론 수업 재현하기

- 이론 수업 끝부분의 titanic data 처리 일부를 직접 재현해보세요.
 - Titanic 데이터셋의 survived, sex, fare, age, embarked 열만 사용하세요.
 - 결측치가 포함된 행, 중복 데이터는 모두 제거하세요.
 - 성별 데이터를 주어진 encoding 함수를 활용해 0과 1로 'sex_code' 열에 표시하세요.
 - 마지막으로, reset_index 메소드를 통해 index를 재설정하세요.
- 처리된 데이터 프레임을 그대로 출력하세요.

2

나이, 탑승 항구 결측치 채우기

- 사용할 데이터가 아래와 같이 주어집니다.
- 주어진 데이터에는 age 및 embarked 데이터에 NaN 데이터가 존재합니다.
- age 데이터 결측치를 승객 나이의 평균값으로 대체하세요!
- Embarked 데이터 결측치는 Embarked 값 중 가장 많이 등장하는 값으로 대체하세요!
- "age"와 "embarked" 컬럼에 대해 value_counts 메소드를 실행한 결과를 차례로 출력하세요.

```
survived  sex   fare  age embarked  sex_code
0         0  male   7.2500  22.0         S         0
1         1 female  71.2833  38.0         C         1
2         1 female   7.9250  26.0         S         1
3         1 female  53.1000  35.0         S         1
4         0  male   8.0500  35.0         S         0
..      ...   ...   ...   ...   ...   ...
886        0  male  13.0000  27.0         S         0
887        1 female  30.0000  19.0         S         1
888        0 female  23.4500   NaN         S         1
889        1  male  30.0000  26.0         C         0
890        0  male   7.7500  32.0         Q         0

[891 rows x 6 columns]
```

주어지는 데이터

데이터분석기초

3

운임 이상치 분석

- 왼쪽 아래 그림과 같이 처리된 데이터가 주어집니다.
- Fare 데이터에 이상치를 분석하고자 합니다. 이상이 있는 데이터는 아래와 같습니다.
 - fare 값이 (제3 사분위수 + 1.5 * (제3 사분위수 - 제1 사분위수))를 초과하는 값은 이상치입니다.
- 이상치인 "fare" 값들의 평균을 출력하세요. 즉, 하나의 숫자가 정답으로 출력됩니다.
- 추가) Colab에서 Seaborns 라이브러리를 이용하여 Fare 데이터의 boxplot을 출력해보세요.

	survived	sex	fare	age	embarked	sex_code
0	0	male	7.2500	22.0	S	0
1	1	female	71.2833	38.0	C	1
2	1	female	7.9250	26.0	S	1
3	1	female	53.1000	35.0	S	1
4	0	male	8.0500	35.0	S	0
..
886	0	male	13.0000	27.0	S	0
887	1	female	30.0000	19.0	S	1
888	0	female	23.4500	28.0	S	1
889	1	male	30.0000	26.0	C	0
890	0	male	7.7500	32.0	Q	0

주어지는 데이터

데이터분석기초

4

성인 여부 표시

- 기본 titanic dataset 에 성인여부 여부를 표시하려고 합니다.
- 1번, 2번 문제로 처리된 데이터가 주어집니다.
- 아래 조건을 만족하는 데이터를 "adult" 컬럼으로 추가하세요.
 - "adult_male"과 "age"를 이용하여 성인임을 판별할 수 있는 기준 나이를 알아내세요.
 - "성인 남성 나이 중 최소 나이"보다 크거나 같다면 성인입니다.
 - 위 기준을 만족하는 모든 사람을 1로, 그렇지 않으면 0으로 "adult" 컬럼에 저장하세요. 저장되는 자료형은 int 형입니다.
- 이후 "adult" 컬럼에 대해 value_counts 메소드를 실행한 결과를 출력하세요.

4

성인 여부 표시

```
> Unnamed: 0 survived sex ... embarked adult_male sex_code
0 0 0 male ... S True 0
1 1 1 female ... C False 1
2 2 1 female ... S False 1
3 3 1 female ... S False 1
4 4 0 male ... S True 0
.. ... ..
765 765 0 female ... Q False 1
766 766 1 female ... S False 1
767 767 0 female ... S False 1
768 768 1 male ... C True 0
769 769 0 male ... Q True 0
```

주어지는 데이터

```
Unnamed: 0 survived sex ... adult_male sex_code adult
0 0 0 male ... True 0 1
1 1 1 female ... False 1 1
2 2 1 female ... False 1 1
3 3 1 female ... False 1 1
4 4 0 male ... True 0 1
.. ... ..
765 765 0 female ... False 1 1
766 766 1 female ... False 1 1
767 767 0 female ... False 1 1
768 768 1 male ... True 0 1
769 769 0 male ... True 0 1
```

변경된 데이터