CNU 데이터 분석 교육

4th lecture

"EDA with Pandas"

2022 - 10 - 24

수학과 최명수

지난시간?

1. Loops : for, while

2. Control Statement : if, else, break

3. Function : Def

오늘은 무엇을?

1. Pandas 패키지를 이용한 EDA 실습 ★ ★ ★

데이터 분석 절차

1. 문제정의
-사회현상, 기업이 마주친 문제를 데이터의 문제로 정의
-데이터 선택

2. 데이터 수집
-선택한 데이터가 잘 확보되어 있는가?

3. 데이터 전처리
-노가다 노가다....
-노이즈 제거, 결측치 처리, 구조변경, 중복제거

4. 데이터 모델링, 마이닝
- 데이터간의 관계를 설정
- 데이터의 패턴을 찾거나 분류 또는 예측

5. 데이터 분석
- 의미있는 결과를 도출하는 과정
- 머신러닝... 예측, 분류 등등

6. 데이터 시각화 및 평가
- 분석결과에 대한 정보를 요약, 압축
- 결과를 해석, 분석 목적과 일치하는지 평가

Pandas

column

	Food	Price	Stock
0	banana	1400	10
1	apple	1000	10
2	soda	900	5
3	chocolate	500	10
4	cookie	1000	10

Column name

index

	Food	Price	Stock
0	banana	1400	10
1	apple	1000	10
2	soda	900	5
3	chocolate	500	10
4	cookie	1000	10

문제 1

다음을 DataFrame으로 작성하여라

	math	sci	eng	final
0	100	90	99	합
1	100	70	97	합
2	90	40	23	불
3	85	93	35	불
4	20	54	56	불
5	50	76	93	불

문제 1

정답

```
# 문제 1

math    = [100,100,90,85,20,50]
science = [90,70,40,93,54,76]
english = [99,97,23,35,56,93]
decision = ["합","합","불","불","불","불"]

pd.DataFrame({'math' : math,
              'sci' : science,
              'eng' : english,
              "final" : decision})
```

Tatanic

★ ★ ★ ★ 분석전에 데이터가 무엇을 설명하는건지 필히 확인해야 한다.

Titanic Data : https://coding-kindergarten.tistory.com/127

컬럼명	의미	인자	자료형(data type)
survived	생존여부	0 (사망) / 1 (생존)	정수(int)
pclass	좌석등급 (숫자)	1 / 2 / 3	정수(int)
sex	성별	male/female	문자열(str)
age	나이	0~80.0	실수(float)
sibsp	형제자매 + 배우자 인원수	0~8	정수(int)
parch:	부모 + 자식 인원수	0~6	정수(int)
fare:	요금	0~512.3292	실수(float)
embarked	탑승 항구	S (Southampton) C (Cherbourg) Q (Queenstown)	문자열(str)
class	좌석등급 (영문)	First, Second, Third	문자열(str)
who	성별	man / woman	문자열(str)
deck	선실 고유 번호 가장 앞자리 알파벳	A,B,C,D,E,F,G	문자열(str)
embark_town	탑승 항구 (영문)	Southampton / Cherbourg / Queenstown	문자열(str)
alive	생존여부 (영문)	no(사망) / yes(생존)	문자열(str)
alone	혼자인지 여부	True (가족 X) / False (가족 O)	참거짓(bool)

Tatanic

컬럼명	의미	인자	자료형(data type)
survived	생존여부	0 (사망) / 1 (생존)	정수(int)
pclass	좌석등급 (숫자)	1 / 2 / 3	정수(int)
sex	성별	male/female	문자열(str)
age	나이	0~80.0	실수(float)
sibsp	형제자매 + 배우자 인원수	0~8	정수(int)
parch:	부모 + 자식 인원수	0~6	정수(int)
fare:	요금	0~512.3292	실수(float)
embarked	탑승 항구	S (Southampton) C (Cherbourg) Q (Queenstown)	문자열(str)
class	좌석등급 (영문)	First, Second, Third	문자열(str)
who	성별	man / woman	문자열(str)
deck	선실 고유 번호 가장 앞자리 알파벳	A,B,C,D,E,F,G	문자열(str)
embark_town	탑승 항구 (영문)	Southampton / Cherbourg / Queenstown	문자열(str)
alive	생존여부 (영문)	no(사망) / yes(생존)	문자열(str)
alone	혼자인지 여부	True (가족 X) / False (가족 O)	참거짓(bool)

lambda

함수 = lambda 변수 : 함수내용

함수(변수) 로 출력

*특정 함수를 만들어서 계속 사용하는 것이 아니라, 아주 잠깐 사용하는 함수일 경우에 사용

ex)

f = lambda x, y : x + y

f(2,3) = 2 + 3 = 5

문제 2

Titanic2 데이터를 이용하여 다음과 같이 n2, n510 index의 age, class, survived를 출력하라

	age	class	survived
n2	26.0	Third	1
n510	29.0	Third	1

문제 2

정답

```
titanic2.loc[['n2','n510'], ['age','class','survived']]
```

	age	class	survived
n2	26.0	Third	1
n510	29.0	Third	1

문제 3

Titanic2 데이터를 이용하여 다음과 같이 처음부터 30번대 데이터까지의 2, 5, 6번째의 column 데이터를 추출하라.

	pclass	sibsp	parch
n0	3	1	0
n1	1	1	0
n2	3	0	0
n3	1	1	0
n4	3	0	0
n5	3	0	0
n6	1	0	0
n7	3	3	1
n8	3	0	2
n9	2	1	0
n10	3	1	1
n11	1	0	0
n12	3	0	0
n13	3	1	5

문제 3

정답

```
titanic2.iloc[:30,[1,4,5]]
```

	pclass	sibsp	parch
n0	3	1	0
n1	1	1	0
n2	3	0	0
n3	1	1	0
n4	3	0	0
n5	3	0	0
n6	1	0	0
n7	3	3	1
n8	3	0	2
n9	2	1	0
n10	3	1	1
n11	1	0	0
n12	3	0	0
n13	3	1	5

문제 4

생존한 승객들의 모든 정보를 추출하라

문제 4

정답

```
titanic[titanic['survived'] == 1]
```

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class	who	adult_male	deck	embark_town	alive	alone
1	1	1	female	38.0	1	0	71.2833	C	First	woman	False	C	Cherbourg	yes	False
2	1	3	female	26.0	0	0	7.9250	S	Third	woman	False	NaN	Southampton	yes	True
3	1	1	female	35.0	1	0	53.1000	S	First	woman	False	C	Southampton	yes	False
8	1	3	female	27.0	0	2	11.1333	S	Third	woman	False	NaN	Southampton	yes	False
9	1	2	female	14.0	1	0	30.0708	C	Second	child	False	NaN	Cherbourg	yes	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
875	1	3	female	15.0	0	0	7.2250	C	Third	child	False	NaN	Cherbourg	yes	True
879	1	1	female	56.0	0	1	83.1583	C	First	woman	False	C	Cherbourg	yes	False
880	1	2	female	25.0	0	1	26.0000	S	Second	woman	False	NaN	Southampton	yes	False
887	1	1	female	19.0	0	0	30.0000	S	First	woman	False	B	Southampton	yes	True
889	1	1	male	26.0	0	0	30.0000	C	First	man	True	C	Cherbourg	yes	True

342 rows × 15 columns

문제 5

여자이거나 20살 미만인 승객의 생존여부와
탑승 클래스를 추출하라

문제 5

정답

```
titanic[(titanic['age'] < 20) | (titanic['sex'] == "female")][['survived', 'class']]
```

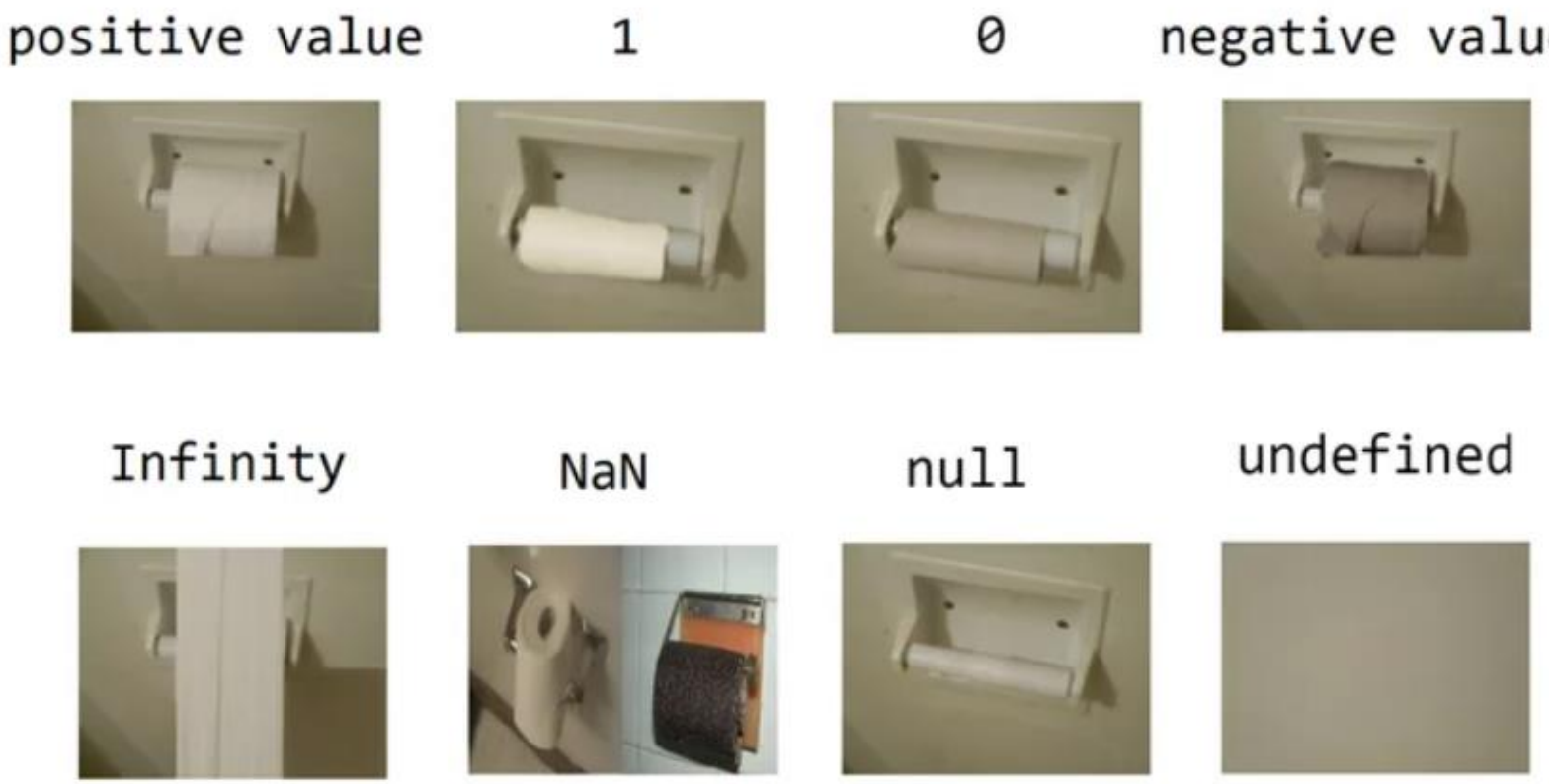
	survived	class
1	1	First
2	1	Third
3	1	First
7	0	Third
8	1	Third
...	...	...
880	1	Second
882	0	Third
885	0	Third
887	1	First
888	0	Third

403 rows × 2 columns

결측치 처리

결측치(Missing Value)
: 값이 없는 것.

*NA(Not Available)
*NaN(Not a Number)
*Null



결측치 처리

1. 데이터 자체를 제거해버린다 (Deletion)

	survived	pclass	sex	age	sibsp	parch	fare	embarked	class
0	0	3	male	22.0	1	0	7.2500	S	Third
3	1	1	female	35.0	1	0	53.1000	S	First
4	0	3	male	35.0	0	0	8.0500	S	Third
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0	13.0000	S	Second
887	1	1	female	19.0	0	0	30.0000	S	First
888	0	3	female	NaN	1	2	23.4500	S	Third
889	1	1	male	26.0	0	0	30.0000	C	First
890	0	3	male	32.0	0	0	7.7500	Q	Third

성별이 뭔지 무슨 수로 판단할건데?

결측치 처리

2. 다른 데이터를 이용하여 채운다(Imputation)

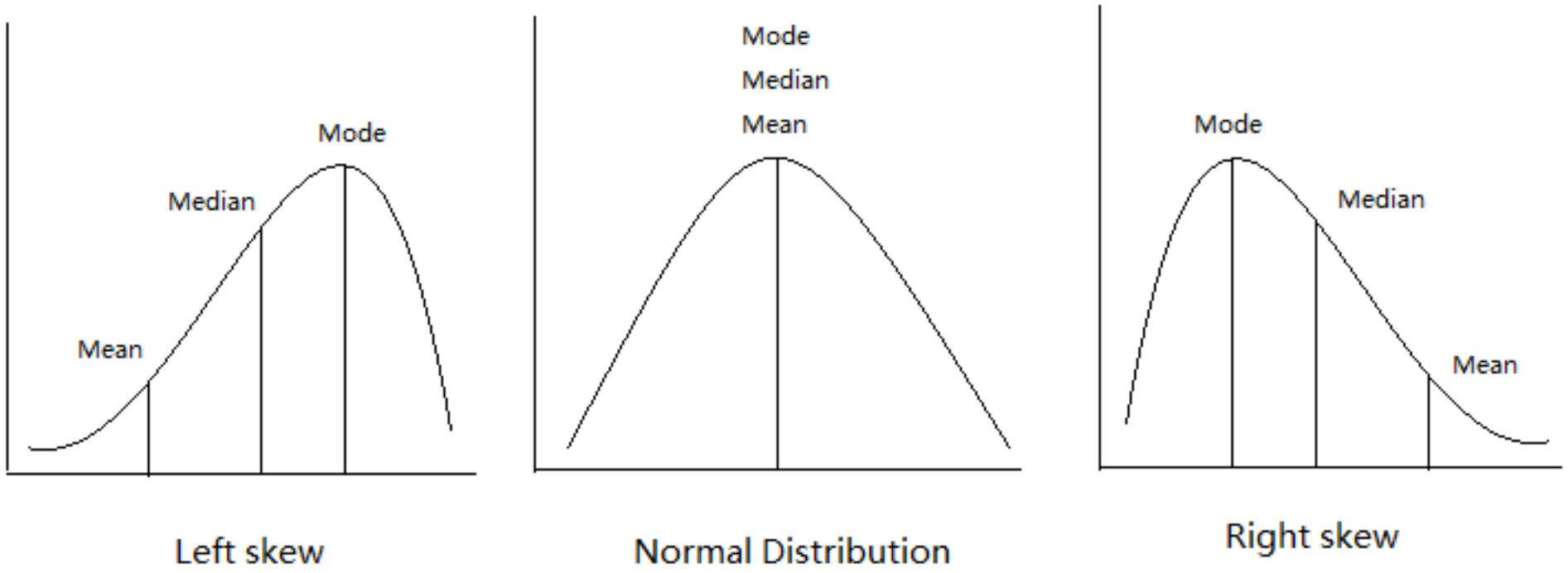
	survived	pclass	sex	age	sibsp	parch	fare	embarked	class
0	0	3	male	22.0	1	0	7.2500	S	Third
1	1	1	female	38.0	1	0	71.2833	C	First
2	1	3	female	26.0	0	0	7.9250	S	Third
3	1	1	female	35.0	1	0	53.1000	S	First
4	0	3	male	35.0	0	0	8.0500	S	Third
...	...	...	...	...	...	...	...	...	...
886	0	2	male	27.0	0	0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

탑승 등급이면 Fare에서 얻을 수 있겠지?

자료의 종류

	종류	예
범주형 자료 (Categorical data)	명목형	성별, 성공여부, 혈액형
	순서형	A>B>C>D, 매우높음>높음>보통>낮음>매우낮음
숫자형 자료 (Numerical data)	이산형	학점, 시험점수, 성공횟수
	연속형	신장, 몸무게, 전구의 수명(시간)

자료의 종류


Mode
Median
Mean
Left skew
Mode
Median
Mean
Normal Distribution
Mode
Median
Mean
Right skew

문제 6

Titanic 데이터를 이용하여 탑승 등급별 (class) 생존자의 비율을 계산하시오

문제 6

힌트 :

일등석 생존자 비율 = 일등석 생존자의 수 / 일등석 생존자 전체 인원
이등석 생존자 비율 = 이등석 생존자의 수 / 이등석 생존자 전체 인원
삼등석 생존자 비율 = 삼등석 생존자의 수 / 삼등석 생존자 전체 인원

문제 6

정답

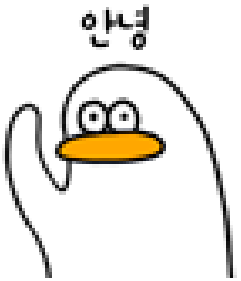
```
titanic.describe(include='all')
titanic['class']


first_all  = titanic[titanic['class'] == 'First' ]['survived'].count()
second_all = titanic[titanic['class'] == 'Second' ]['survived'].count()
third_all  = titanic[titanic['class'] == 'Third' ]['survived'].count()

first_survived  = titanic[(titanic['class'] == 'First') & (titanic['survived'] == 1)]['survived'].count()
second_survived = titanic[(titanic['class'] == 'Second') & (titanic['survived'] == 1)]['survived'].count()
third_survived  = titanic[(titanic['class'] == 'Third') & (titanic['survived'] == 1)]['survived'].count()

first_ratio  = first_survived/ first_all
second_ratio = second_survived / second_all
third_ratio  = third_survived / third_all

print("일등석 생존자 비율 = {}".format(first_ratio))
print("이등석 생존자 비율 = {}".format(second_ratio))
print("삼등석 생존자 비율 = {}".format(third_ratio))
```

```
일등석 생존자 비율 = 0.6296296296296297
이등석 생존자 비율 = 0.4728260869565217
삼등석 생존자 비율 = 0.24236252545824846
```

공지

1. 시험 잘 보 세 오

끝



담에봅시당