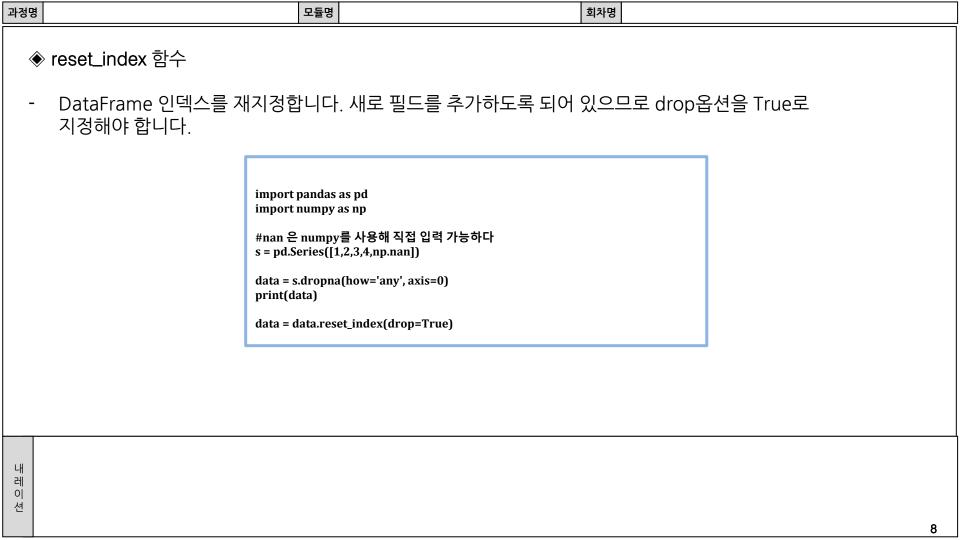
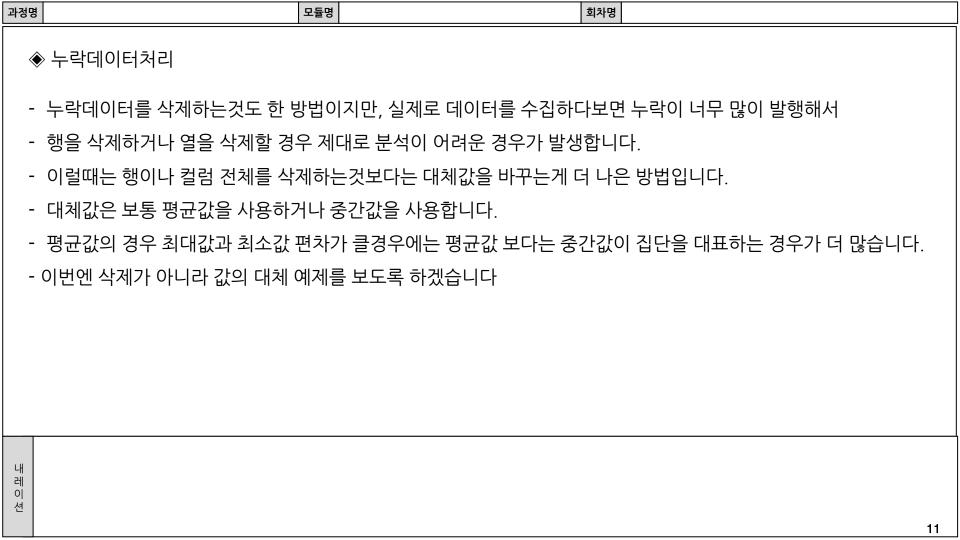


과정명	모듈명 회차명
•	dropna 함수
-	DataFrame.dropna(self, axis=0, how='any', thresh=None, subset=None, inplace=False)[source]
-	subset에서 지정한 필드들에 NaN값이 있는 행을 삭제한다. inplace=True로 지정하면 자신의 데이터가 수정된다. 필드 지정방법은 list형태이다. subset=['필드1', '필드2']
-	thresh : int값 이 필드에서 부여한 개수만큼 NaN이 없는 행이나 열을 삭제
-	how : any, all any일 경우 행이나 컬럼값중에 하나라도 있으면 행이나 열을 삭제, all의 경우 행이나 컬럼의 값들이 모두 NaN 이 있을때 행이나 컬럼을 삭제합니다.
-	axis : 0은 행 1은 열을 지칭. 삭제할 방향을 지시한다
내 레 이 션	7



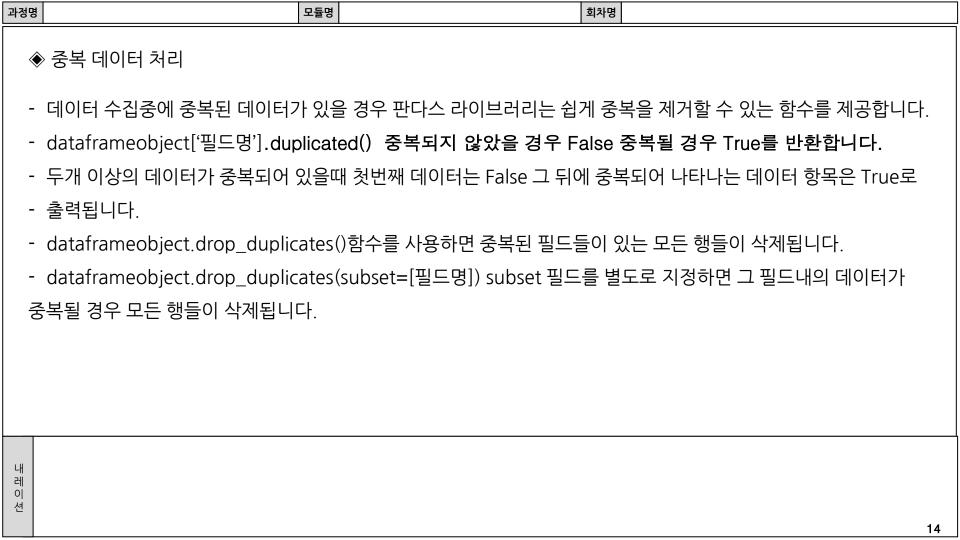
과정명	!	모듈명	회차명	
•	▶ 누락데이터처리			
#	파일명 : exam11_1.py			
ir	nport pandas as pd			
	header가 3번째 줄에 있음 ata = pd.read_csv("./data/data.csv")			
p p	rint("컬럼명 : ", data.columns) rint("인덱스 : ", data.index)			
р	rint( data.info() )			
p p	누란된 데이터가 있는지 확인하자. rint( data['height'].value_counts(dropna=False) rint( data['height'].isnull()) #NaN이면 True, 아니 rint( data['height'].notnull())	) I면 False반환		
	null값인것에 대한 개수 확인 rint( data['height'].isnull().sum(axis=0))			
내 레				
이 션				
				9

과정명		모듈명	회차명
<b>©</b>	▶ 누락데이터처리		
	nan값을 thresh개 가진 열 모두 삭제 ata_thresh = data.dropna(axis=1, thresh=3)		
	합계 int ( data['height'].sum() )		
pr da pr	rint("데이터 개수") rint(data.shape) ata = data.dropna(subset=['height'], how='any' rint("삭제후 데이터개수") rint(data.shape)	, axis=0)	
pr	ata = data.dropna(subset=['weight'], how='any' rint("삭제후 데이터개수") rint(data.shape)	', axis=0)	
da	인덱스 다시 부여 ata = data.reset_index(drop=True) rint(data)		
내			
레 이 션			
			10



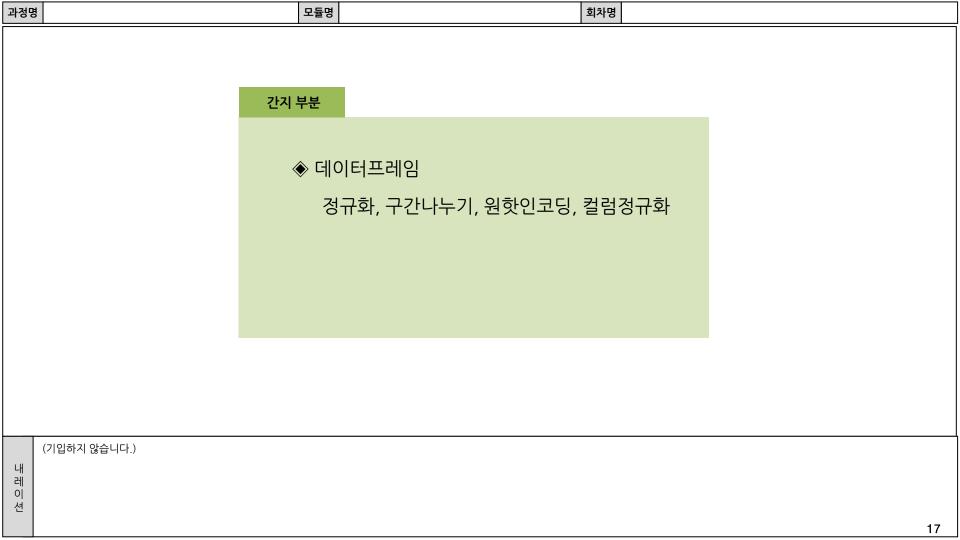
과정명		모듈명		회차명	
•	· 누락데이터처리				
	ㅏ일명∶exam11_2.py -락데이터 치환하기				
im	port pandas as pd				
	eader가 3번째 줄에 있음 ta = pd.read_csv("./data/data.csv")				
pri pri	=란된 데이터가 있는지 확인하자. nt( data['height'].value_counts(dropna=False)) nt( data['height'].isnull()) #NaN이면 True, 아니 nt( data['height'].notnull())		se반환		
pri	ull값인것에 대한 개수 확인 nt("height 필드 Nan 개수 : ", data['height'].isn nt("weight 필드 Nan 개수 : ", data['weight'].isr				
#ት me	nt("누락값 대체 후") =락값 대체하기 ean_height = data['height'].mean() ean_weight = data['weight'].mean()				
내 레 이					
션					1

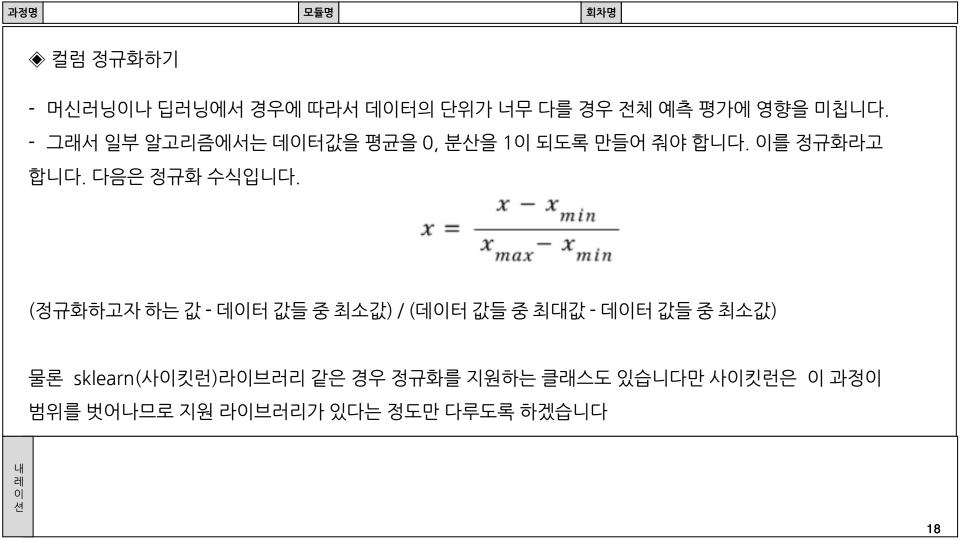
과정명	모듈명	회차명
•	» 누락데이터처리	weight 필드 Nan 개수 : 2 누락값 대체 후
da	ata['height'].fillna( mean_height, inplace=True) ata['weight'].fillna( mean_weight, inplace=True)	name height weight 0 A1 180.000000 92.000000 1 A2 176.000000 70.000000 2 A3 175.000000 65.000000 3 A4 172.000000 64.000000
pri pri	int("height 필드 Nan 개수 : ", data['height'].isnull().sum(axis int("weight 필드 Nan 개수 : ", data['weight'].isnull().sum(axi	s=0)) 5 A6 175.000000 74.000000 68.000000 6 A7 177.000000 68.000000
pri	int( data )	7 A8 172.000000 65.000000 8 A9 177.888889 77.000000 9 A10 173.000000 59.000000 10 A11 174.000000 66.000000
	illna 함수 NaN 값을 특정 값으로 대체 합니다. nplace옵션을 True를 해줘야 값이 바뀝니다	11 A12 177.000000 63.000000 12 A13 179.000000 65.000000 13 A14 181.000000 72.000000 14 A15 189.000000 79.000000 15 A16 193.000000 84.000000 16 A17 177.888889 34.000000 17 A18 172.000000 160.000000 18 A19 179.000000 84.000000 19 A20 169.00000 64.000000 20 A21 167.000000 56.000000 21 A22 230.000000 12.000000 22 A23 171.000000 15.000000 23 A24 177.888889 80.000000 24 A25 187.000000 78.000000 25 A26 176.000000 67.107143 26 A27 175.000000 59.000000 27 A28 173.000000 59.000000 28 A29 169.000000 53.000000 29 A30 174.000000 59.000000
내 레 이 션		

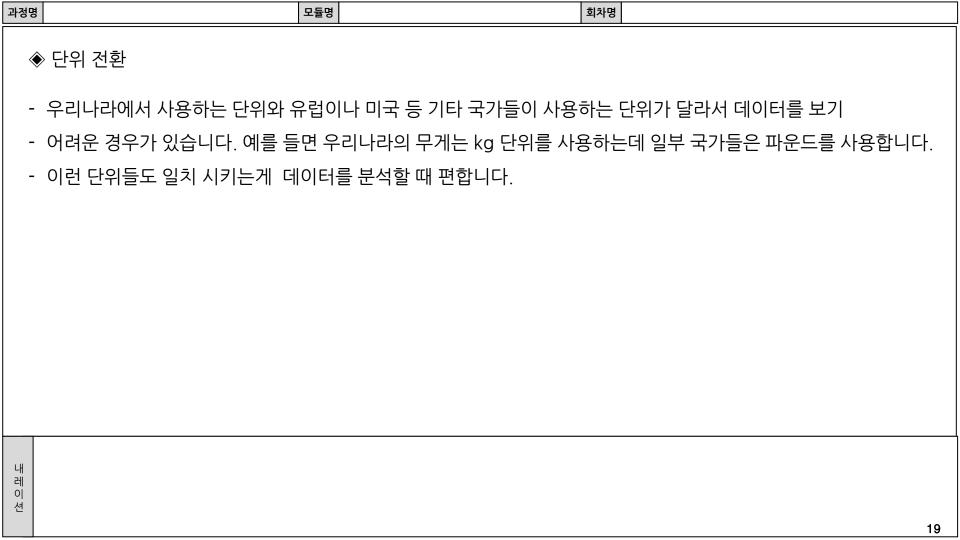


과정명	ļ	모듈명		회차명	
«	▶ 누락데이터처리				
	파일명 : exam11_3.py 중복데이터 제거하기				
ir	nport pandas as pd				
}	ata = { 'passenger_code':['A101', 'A102', 'A103', 'A1 'target':['광주', '서울', '부산', '광주', '대구', '공 'price':[25000, 27000, 45000, 25000, 35000,	}주', '투	<del>'</del> 산'],		
c	f = pd.DataFrame(data)				
p	rint( df )				
С	rint("중복된 데이터 ") ol = df['passenger_code'].duplicated() #중복된 rint( col )	데이터	확인하기		
내 레 이 :					
션				15	

과정명		모듈명	회차명
•	› 누락데이터처리		
df2	등복 제거시 모든 데이터가 일치된것만 제거한다. 2 = df.drop_duplicates() nt( df2 )		
df:	nt("특정 컬럼값이 중복일때 제거하기") 3 = df.drop_duplicates(subset=['passenger_cod nt( df3 )	de'])	
df	nt("두개의 컬럼값이 중복일때 제거하기") 3 = df.drop_duplicates(subset=['passenger_coo nt( df3 )	de', 'target'])	
내 레 이			
션			16







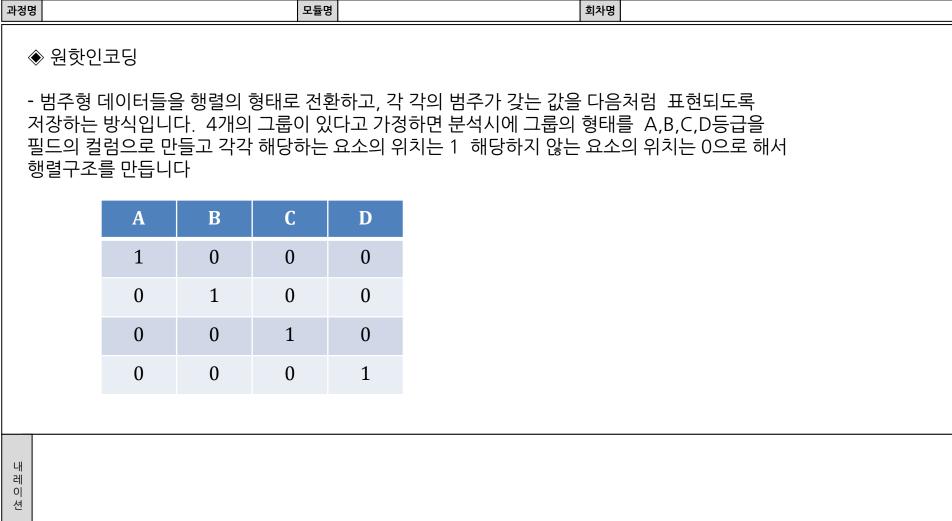
과정명	모듈명	회차명	<b>,</b>	
<b>(</b>	> 정규화및 단위환산			
	파일명 : exam11_4.py 데이터표준화			
im	port pandas as pd			
l pr	ata = pd.read_csv('./data/auto-mpg.csv') rint(data.info()) rint(data.head())			
da	컬럼명 변경하기 ata.columns=['mpg', 'cyl', 'disp', 'power', 'weight', 'acce', 'model']	Г		
#- #( da	rint(data.head()) 정규화 (정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 값들 중 최대값 - 데이터 값들 중 최소값) ata['mpg2'] = (data['mpg'] - data['mpg'].min())/(data['mpg'].max()-data['mpg'].min()) rint(data)		정규화는 데이터분석에서 아주 중요한 작업입니다. 정규화 작업을 해줘야만 하는 머신러닝 알고리즘이 많이 있습니다.	
	단위환산 - 한국 단위로 환산하기 pg_unit = 1.60934 / 3.78541			
	ata['kpl'] = (data['mpg'] * mpg_unit).round(2) int( data.head() )			
내				
레이				
션				20

과정명	모듈명 회차명	
•	타입전환	
-	파이썬은 숫자형태의 데이터는 실제 데이터가 어떤 형태로 저장되던(엑셀 같은경우 숫자를 문자열형태로 저장	
-	하기도 하고 숫자 형태로 저장하기도 합니다.) 숫자자료로 잘 읽어와서 수치 자료는 크게 문제가 되지 않습니다	
-	문제가 되는 경우는 수치자료가 와야 할 열에 수치 자료가 아닌 다른 문자가 가끔씩 끼어 들어올 때 이에 대한	
	처리를 해줘야 합니다. 문자를 제거하거나 문자가 들어있는 행의 데이터를 버리거나 하고 유형을 전환해줘야	
	합니다.	
-	문자열의 경우에는 분석시에 문자열 자체로 사용을 못합니다. 문자열의 경우에는 범주형 자료 타입으로 전환을	
	해야 합니다.	
-	이번에는 타입 전환에 대한 예에 대해서 살펴보도록 하겠습니다.	
-	타입전환은 .astype 함수를 이용해서 합니다	
내		
레 이 션		
		21

과정명		모듈명	회차명		
<b>«</b>	<b>▶</b> 타입전환				
	파일명 : exam11_5.py 데이터표준화				
	nport pandas as pd nport numpy as np				
pr	ata = pd.read_csv('./data/auto-mpg.csv') rint(data.info()) rint(data.head())				
## #3 da pr	타입이 맞지 않을 경우 전환을 해서 사용해야 한다 현재 사용하는 파이썬 버전은 문자열 데이터라도 <sup>=</sup> 파이썬 버전에 따라 다르게 동작할 수 도 있다 ata.columns=['mpg', 'cyl', 'disp', 'power', 'weig int(data.dtypes) rint(data.head())	<b>수치 형태면 자동으로</b>			
pr	rint( data['disp'].unique())				
내 레 이 션					
				20	2

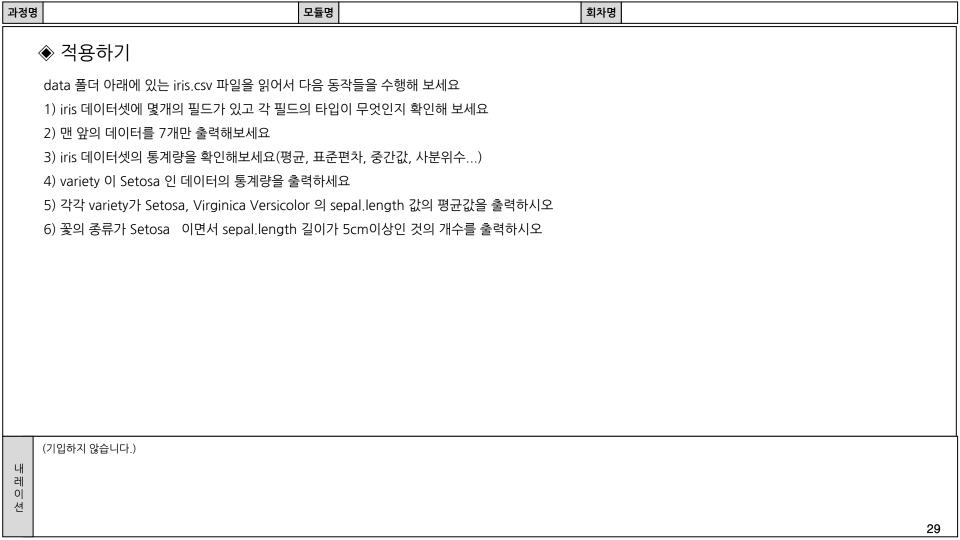
과정명	모듈명	회차명
	◈ 타입전환	
	#잘못된 데이터를 NaN으로 먼저 바꾼다 data['disp'].replace('?', np.nan, inplace=True)	
	<pre>print(data.head())  data.dropna(subset=['disp'], axis=0, inplace=True) print(data.head())</pre>	데이터를 범주형으로 바꿀때 astype('category')를 사용합니다. 문자열의 형태로 입력된 데이터는 분석시에 꼭 범주형으로 바꾸어서 처리해야 합니다.
	<pre>print(data.dtypes) data['disp'] = data['disp'].astype('float') print(data.dtypes)</pre>	
	#범주형으로 바꾼다 data['model'] = data['model'].astype('category') print(data.dtypes)	
내		
레 이 션		
		23

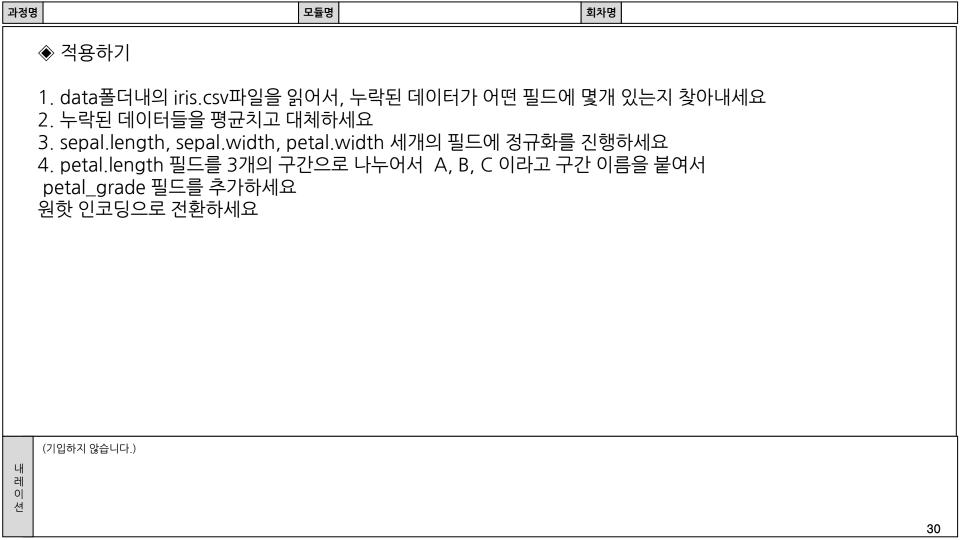
과정명	모듈명 회차명	
•	» 구간나누기	
-	연속적 데이터라고 하더라도 경우에 따라서는 불연속 데이터로 전환해야 할때가 있습니다.	
-	예를들어서 연비의 경우 국가에서 특정 조건을 만족할 경우 구간을 나누어서 등급을 부여할 수있습니다.	
-	연비를 수치자료로 듣는 것 보다 이 자동차의 연비는 1등급이야 또는 2등급이야 라고 할때 실제 수치를	
-	듣는것보다 정보가 빠르고, 연비 등급별로 각 모델들에 대한 평가나 분석을 해야 하는 경우도 있습니다.	
-	이때 구간나누기를 통해 연속 데이터를 불연속 데이터로 변환이 가능합니다.	
-	불연속 데이터로 전환 후 실제 분석시에는 특별한 방법을 사용하여 데이터를 저장해야 합니다. 만일 연비를	
	등급으로 A,B,C,D 4개로 나누었다면, 머신러닝이나 딥러닝 분석에서 이를 그대로 사용하지 못합니다.	
	머신러닝이나 딥러닝은 수학입니다. 저 등급은 다음과 같은 원핫인코딩 형태로 전환해야 사용이	
	가능합니다.	
내 레 이		
션		24



과정명		모듈명		회차명	B	
# d # # p d C p b d	H주형으로 바꾼다 ata['model'] = data['model']. 연속적인 값을 구간으로 나누어 print(data['power'])  rint(data['power'].isnull().sum ata.dropna(subset=['power'], ount, bin_dividers = np.histogr rint( bin_dividers)  in_names = ["D", "C", "B", "A ata["grade"] = pd.cut( x=data labels = bin_names, include_ rint( data )	비연 (axis inpla am(c	속적인값(범주형)으로 나눈다 =0)) lce=True) data['power'], bins=4)		np.histogram 은 구간을 나누는 함수이다. 데이터의 범위를 알아내서 지정된 구간을 나누고 구간의 경계에 대한 값을 가져온다. 구간을 나누고자 하는 필드와 구간의 개수를 전달하면 구간의 경계, 구간의 데이터개수를 반환한다.  cut함수, x인자에 필드를 정해주고, bins 인자에 np.histogram이 생성한 구간 정보를 전달한다. labels 에 각 구간에 붙일 라벨명을 지정하고 include_lowest=True는 경계값을 포함하라는 의미임	
내 레 이 션						27

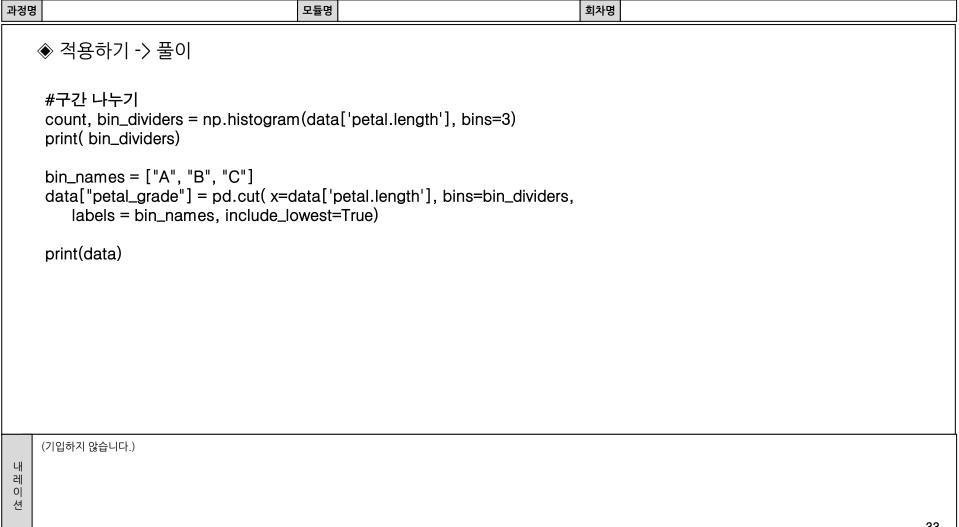
과정명	모듈명	회차명
# Y. fr ei ei Y.	아마ehot encoding _class = np.array(data['grade']).reshape(-1,1)  om sklearn.preprocessing import OneHotEncoder nc = OneHotEncoder() nc.fit(Y_class)  _class_onehot = enc.transform(Y_class).toarray() _class_recovery = np.argmax(Y_class_onehot, axis=1).reshape(-1,1)  rint(Y_class_onehot)	onehot 인코딩을 진행라면 데이터 필드가 1차원인 상태로는 진행이 되지 않습니다. reshape함수를 이용해서 차원을 먼저 바꿔주고 sklearn.preprocessing 의 OneHotEncoder객체를 만든 후 transform함수를 수행하면됩니다. onehot encoding 을 복원하려면 np.argmax 함수를 사용하면 됩니다. no.argmax함수는 전달 받은 함수중에서 가장 큰값을 갖는 인자를 반환하는 함수입니다.
내 레 이 션		28





과정명		모듈명	회차명
	◈ 적용하기- 풀이		
	import pandas as pd import numpy as np		
-	data = pd.read_csv('./data/iris.csv' #NaN값 있는지 체크하기 - 각 필드별 print( data.isnull().sum())		
;	sepal_length_mean = data['sepal.le sepal_width_mean = data['sepal.wi petal_length_mean = data['petal.le petal_width_mean = data['petal.wic	idth'].mean() ngth'].mean()	
	data['sepal.length'].fillna( sepal_le data['sepal.width'].fillna( sepal_wid data['petal.length'].fillna( petal_len data['petal.width'].fillna( petal_wid	dth_mean, inplace=True) ngth_mean, inplace=True)	
내레	(기입하지 않습니다.)		
이 년			
			31

과정명		모듈명	회차명	
<	◆ 적용하기 -> 풀이			
	#정규화를 함수로 만들었음 def normalize(columnname): max = data[columnname].max min = data[columnname].min( return (data[columnname]-min	)		
	data['sepal.length'] = normalize('s data['sepal.width'] = normalize('s data['petal.width'] = normalize('pe	epal.width')		
	print(data)			
내 레 이	(기입하지 않습니다.)			
션			32	



과정명		모듈명	회차명
<	◈ 적용하기 -> 풀이		
	#onehot 인코딩 Y_class = np.array(data['petal_gra	ade']).reshape(-1,1)	
	from sklearn.preprocessing import enc = OneHotEncoder() enc.fit(Y_class)	: OneHotEncoder	
	Y_class_onehot = enc.transform(Y_class_recovery = np.argmax(Y_c	′_class).toarray() class_onehot, axis=1).reshape(-1,1)	
	print(Y_class_onehot)		
내레	(기입하지 않습니다.)		
이 이 선			34

과정명	성명 모듈명 회차명	
•	◈ 문제풀기 문제 1) 다음 설명중 잘못 된 것은?	
	<ul> <li>① 데이터의 누락치는 데이터 분석에 지대한 영향을 미치므로 모두 삭제해야 한다</li> <li>② 데이터 누락치는 데이터 분석에 지대한 영향을 미치므로 평균값이나 중간값등으로 대체해야 한다</li> <li>③ 데이터의 정규화는 분석의 정확도를 위해 필요한 경우가 많다</li> <li>④ 데이터의 타입이 문자열일 경우 범주형자료로 전환해야 한다</li> </ul>	
	정답) 5 해설) 누락된 데이터가 너무 많을 경우 모두 삭제는 오히려 분석 결과에 안좋은 영향을 미칠 수 있으므로 평균치나 중간값으로 대체 하는 방법을 취하는게 좋습니다. 난이도) 1(1:아주어려움, 2:어려움 3: 보통 4: 쉬움 5: 아주 쉬움) 관련페이지번호) 9	
내 레 이 션		35

번 호	문제	정답	난이도	해설	관련학습보기
2	다음 데이터에서 누락 데이터를 찾고자 합니다. 함수를 바르게 적용한 경우는 ? ① data['height'].isnull().sum() ② data['height'].NaNsum() ③ data['height'].CountNan()	1	4	isnull() 함수와 sum() 함수를 호출하다	7
	④ data['height'].CntNa()				
3	데이터중 누락된 데이터(NaN)을 삭제하려고 할 경우 적당한 함수는? ① isnull ② dropna ③ deleterow ④ removena	2	5	dropna함수를 사용한다	9
4	데이터를 삭제하면 인덱스 번호가 뒤죽박죽이 됩니다. 삭제후 인덱스를 재부여 하려면 어떤 함수를 써야 하나요? ① set_index ② refresh_index ③ reset_index ④ index	3	4	reset_index 함수를 사용하면 index가 새로 부여된다.	10

번 호	문제	정답	난이 도	해설	관련학습 보기
5	잘못된 데이터가 들어온경우 처리법으로 가장 적절한 것은         ① 데이터를 메모장에서 직접 수정한다         ② 예외처리를 한다         ③ 오류 데이터를 NaN값으로 모두 전환후 대체값으로 대신하던가 삭제한다         ④ 오류데이터를 엑셀 기능을 이용해 모두 삭제한다.	3	3	데이터 건수가 많을 경우에는 일일이 메모장이나 엑셀에서 작업 이 어렵습니다. 우선 데이터를 읽어서 NaN값으로 모두 변환후 한번에 처리하는게 편합니다.	13
6	다음궁 정규화 수식으로 올바른것은?  ① (정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 값들 중 최대 값 - 데이터 값들 중 최소값) / (데이터 과들 중 최소값)  ② (정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 평균값)  ③ (정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 분산)  ④ (정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 표준편차)	1	1	(정규화하고자 하는 값 - 데이터 값들 중 최소값) / (데이터 값들 중 최대값 - 데이터 값들 중 최소값)	22
7	다음 등급을 원핫인코딩 방식으로 변환하려고 합니다. 올바른 변환을 기술 하십시요 (원핫인코딩 한 상태의 값을 기술하세요) 등급: 좋음, 보통, 나쁨	해설참조	2	좋음 보통 나쁨 1 0 0 0 1 0 0 0 1	27

번호	문제	정답	난이 도	해설	관련학습보기
8	다음 궁 구간 나누기를 하는 바른 사유는         ① 연속형 값을 비연속형 값의 형태로 전환하려고         ② 구간별 평균값을 구하려고         ③ 히스토그램 차트를 그리기 위해서         ④ 비연속형 값을 연속형 값으로 전환하려고	1	5	연속형값을 비연속형으로 전환할 필요가 있어서	26
9	dropna 함수를 사용해 데이터를 삭제하고자 할때 inplace=True 옵션이 하는일은 ? 예)data.dropna(subset=['disp'], axis=0, inplace=True)	자신의 데이터를 수강한다	1	inplace속성을 True 로 해야 삭제된 데이터가 자신의 데이터로 적용됩니다.	9
10	다음 데이터셋중에서 필드 mpg와 cylinders 두 개의 필드에서 데이터가 NaN값이 있으면 삭제하려고 합니다. 다음 코드의 나머지 부분을 완성하세 요 data.dropna( , axis=0, inplace=True)	subset= ['mpg', 'cyliend ers']	1	subset를 지정해주면 됩니다. subset는 리스트 형태로 전돨됩니다	18

