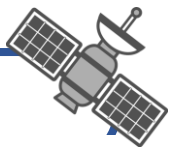
https://github.com/multicampus-prods/test-practicing-associateds



Dataset_01.csv

구분자 : comma(“,”), 4,572 Rows, 5 Columns, UTF-8 인코딩

글로벌 전자제품 제조회사에서 효과적인 마케팅 방법을 찾기 위해서 채널별 마케팅 예산과 매출금액과의 관계를 분석하고자 한다.

컬 럼	정 의	Type
TV	TV 마케팅 예산 (억원)	Double
Radio	라디오 마케팅 예산 (억원)	Double
Social_Media	소셜미디어 마케팅 예산 (억원)	Double
Influencer	인플루언서 마케팅 (인플루언서의 영향력 크기에 따라 Mega / Macro / Micro / Nano)	String
SALES	매출액	Double

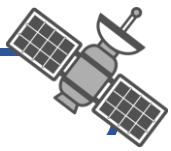
1. 데이터 세트 내에 총 결측값의 개수는 몇 개인가? **(답안 예시) 23**

2. TV, Radio, Social Media 등 세 가지 다른 마케팅 채널의 예산과 매출액과의 상관분석을 통하여 각 채널이 매출에 어느 정도 연관이 있는지 알아보고자 한다.
   - 매출액과 가장 강한 상관관계를 가지고 있는 채널의 상관계수를 소수점 5번째 자리에서 반올림하여 소수점 넷째 자리까지 기술하시오. **(답안 예시) 0.1234**

3. 매출액을 종속변수, TV, Radio, Social Media의 예산을 독립변수로 하여 회귀분석을 수행하였을 때, 세 개의 독립변수의 회귀계수를 큰 것에서부터 작은 것 순으로 기술하시오.
   - 회귀계수는 소수점 넷째 자리 이하는 버리고 소수점 셋째 자리까지 기술하시오. **(답안 예시) 0.123**

Python	import pandas as pd from sklearn.linear_model import LinearRegression

https://github.com/multicampus-prods/test-practicing-associateds



Dataset_02.csv

구분자 : comma(“,”), 200 Rows, 6 Columns, UTF-8 인코딩

환자의 상태와 그에 따라 처방된 약에 대한 정보를 분석하고자 한다.

컬럼	정의	Type
Age	연령	Integer
Sex	성별	String
BP	혈압 레벨	String
Cholesterol	콜레스테롤 레벨	String
Na_to_k	혈액 내 칼륨에 대비한 나트륨 비율	Double
Drug	Drug Type	String

1. 해당 데이터에 대한 EDA를 수행하고, 여성으로 혈압이 High, Cholesterol이 Normal인 환자의 전체에 대비한 비율이 얼마인지 소수점 네 번째 자리에서 반올림하여 소수점 셋째 자리까지 기술하시오. **(답안 예시) 0.123**

2. Age, Sex, BP, Cholesterol 및 Na_to_k 값이 Drug 타입에 영향을 미치는지 확인하기 위하여 아래와 같이 데이터를 변환하고 분석을 수행하시오.
   - Age_gr 컬럼을 만들고, Age가 20 미만은 ‘10’, 20부터 30 미만은 ‘20’, 30부터 40 미만은 ‘30’, 40부터 50 미만은 ‘40’, 50부터 60 미만은 ‘50’, 60이상은 ‘60’으로 변환하시오.
   - Na_K_gr 컬럼을 만들고 Na_to_k 값이 10이하는 ‘Lv1’, 20이하는 ‘Lv2’, 30이하는 ‘Lv3’, 30초과는 ‘Lv4’로 변환하시오.
   - Sex, BP, Cholesterol, Age_gr, Na_K_gr이 Drug 변수와 영향이 있는지 독립성 검정을 수행하시오.
   - 검정 수행 결과, Drug 타입과 연관성이 있는 변수는 몇 개인가? 연관성이 있는 변수 가운데 가장 큰 p-value를 찾아 소수점 여섯 번째 자리 이하는 버리고 소수점 다섯 번째 자리까지 기술하시오.

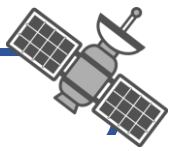
   (답안 예시) 3, 1.23456

3. Sex, BP, Cholesterol 등 세 개의 변수를 다음과 같이 변환하고 의사결정나무를 이용한 분석을 수행하시오.
   - Sex는 M을 0, F를 1로 변환하여 Sex_cd 변수 생성
   - BP는 LOW는 0, NORMAL은 1 그리고 HIGH는 2로 변환하여 BP_cd 변수 생성
   - Cholesterol은 NORMAL은 0, HIGH는 1로 변환하여 Ch_cd 생성
   - Age, Na_to_k, Sex_cd, BP_cd, Ch_cd를 Feature로, Drug을 Label로 하여 의사결정나무를 수행하고 Root Node의 split feature와 split value를 기술하시오.

이 때 split value는 소수점 셋째 자리까지 반올림하여 기술하시오. **(답안 예시) Age, 12.345**

Python	#2 import scipy.stats as stats #3 from sklearn.tree import DecisionTreeClassifier from sklearn.tree import export_graphviz import pydot import graphviz

https://github.com/multicampus-prods/test-practicing-associateds



Dataset_04.csv

구분자 : comma(“,”), 6,718 Rows, 4 Columns, UTF-8 인코딩

한국인의 식생활 변화가 건강에 미치는 영향을 분석하기에 앞서 육류 소비량에 대한 분석을 하려고 한다. 확보한 데이터는 세계 각국의 1인당 육류 소비량 데이터로 아래와 같은 내용을 담고 있다.

컬 럼	정 의	Type
LOCATION	국가명	String
SUBJECT	육류 종류 (BEEF / PIG / POULTRY / SHEEP)	String
TIME	연도 (1990 ~ 2026)	Integer
Value	1인당 육류 소비량 (KG)	Double

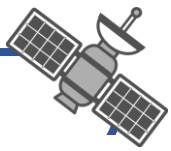
1. 한국인의 1인당 육류 소비량이 해가 갈수록 증가하는 것으로 보여 상관분석을 통하여 확인하려고 한다.
   - 데이터 파일로부터 한국 데이터만 추출한다. 한국은 KOR로 표기되어 있다.
   - 년도별 육류 소비량 합계를 구하여 TIME과 Value간의 상관분석을 수행하고 상관계수를 소수점 셋째 자리에서 반올림하여 소수점 둘째 자리까지만 기술하시오. **(답안 예시) 0.55**

2. 한국 인근 국가 가운데 식생의 유사성이 상대적으로 높은 일본(JPN)과 비교하여, 연도별 소비량에 평균 차이가 있는지 분석하고자 한다.
   - 두 국가의 육류별 소비량을 연도기준으로 비교하는 대응표본 t 검정을 수행하시오.
   - 두 국가 간의 연도별 소비량 차이가 없는 것으로 판단할 수 있는 육류 종류를 모두 적으시오. (알파벳 순서) **(답안 예시) BEEF, PIG, POULTRY, SHEEP**

3. (한국만 포함한 데이터에서) Time을 독립변수로, Value를 종속변수로 하여 육류 종류(SUBJECT)별로 회귀분석을 수행하였을 때, 가장 높은 결정계수를 가진 모델의 학습오차 중 MAPE를 반올림하여 소수점 둘째 자리까지 기술하시오. **(답안 예시) 21.12**
   (MAPE : Mean Absolute Percentage Error, 평균 절대 백분율 오차)(MAPE = Σ (| y − y^ | / y) * 100/n))

Python	#1 import pandas as pd import numpy as np #2 from scipy.stats import ttest_rel #3 from sklearn.linear_model import LinearRegression

https://github.com/multicampus-prods/test-practicing-associateds



원룸촌에 위치한 A마트는 데이터 분석을 통해 보다 체계적인 재고관리와 운영을 하고자 한다. 이를 위해 다음의 두 데이터 세트를 준비하였다.

Dataset_15_Mart_POS.csv

구분자 : comma(","), 20488 Rows, 3 Columns, UTF-8 인코딩

컬럼	정의	Type
Member_number	고객 고유 번호	Double
Date	구매일	String
itemDescription	상품명	String

Dataset_15_item_list.csv

구분자 : comma(","), 167 Rows, 4 Columns, UTF-8 인코딩

컬럼	정의	Type
prod_id	상품 고유 번호	Double
prod_nm	상품명	String
alcohol	주류 상품 여부(1 : 주류)	Integer
frozen	냉동 상품 여부(1 : 냉동)	Integer

1. (Dataset_15_Mart_POS.csv를 활용하여) 가장 많은 제품이 팔린 날짜에 가장 많이 팔린 제품의 판매 개수는? **(답안 예시) 1**

2. (Dataset_15_Mart_POS.csv, Dataset_15_item_list.csv를 활용하여) 고객이 주류 제품을 구매하는 요일이 다른 요일에 비해 금요일과 토요일이 많을 것이라는 가설을 세웠다. 이를 확인하기 위해 금요일과 토요일의 일별 주류제품 구매 제품 수 평균과 다른 요일의 일별 주류제품 구매 제품 수 평균이 서로 다른지 비교하기 위해 독립 2표본 t검정을 실시하시오.
해당 검정의 p-value를 기술하시오.
- 1분기(1월 ~ 3월) 데이터만 사용하여 분석을 실시하시오.
- 등분산 가정을 만족하지 않는다는 조건 하에 분석을 실시하시오.
- p-value는 반올림하여 소수점 둘째 자리까지 기술하시오. **(답안 예시) 0.12**

3. (Dataset_15_Mart_POS.csv를 활용하여) 1년 동안 가장 많이 판매된 10개 상품을 주력 상품으로 설정하고 특정 요일에 프로모션을 진행할지 말지 결정하고자 한다. 먼저 요일을 선정하기 전에 일원 분산 분석을 통하여 요일별 주력 상품의 판매 개수의 평균이 유의미하게 차이가 나는지 알아보고자 한다. 이와 관련하여 일원 분산 분석을 실시하고 p-value를 기술하시오.
- p-value는 반올림하여 소수점 둘째 자리까지 기술하시오. **(답안 예시) 0.12**

Python	from statsmodels.formula.api import ols from statsmodels.stats.anova import anova_lm

정답

세트	번호	답안
SET 01	Q1	
	Q2	
	Q3	
SET 02	Q1	
	Q2	
	Q3	
SET 03	Q1	
	Q2	
	Q3	
SET 04	Q1	
	Q2	
	Q3	
SET 05	Q1	
	Q2	
	Q3	
SET 06	Q1	
	Q2	
	Q3	
SET 07	Q1	
	Q2	
	Q3	
SET 08	Q1	
	Q2	
	Q3	

세트	번호	답안
SET 09	Q1	
	Q2	
	Q3	
SET 10	Q1	
	Q2	
	Q3	
SET 11	Q1	
	Q2	
	Q3	
SET 12	Q1	
	Q2	
	Q3	
SET 13	Q1	
	Q2	
	Q3	
SET 14	Q1	
	Q2	
	Q3	
SET 15	Q1	
	Q2	
	Q3	