#### 수원대 & 데이콘

# AI 경진대회 기반일

Data Science 실무 역량 강화 교육



정성문

DACON / DS / MANAGER 산업 데이터 공학 전공 농산물 가격예측 경진대회 수상

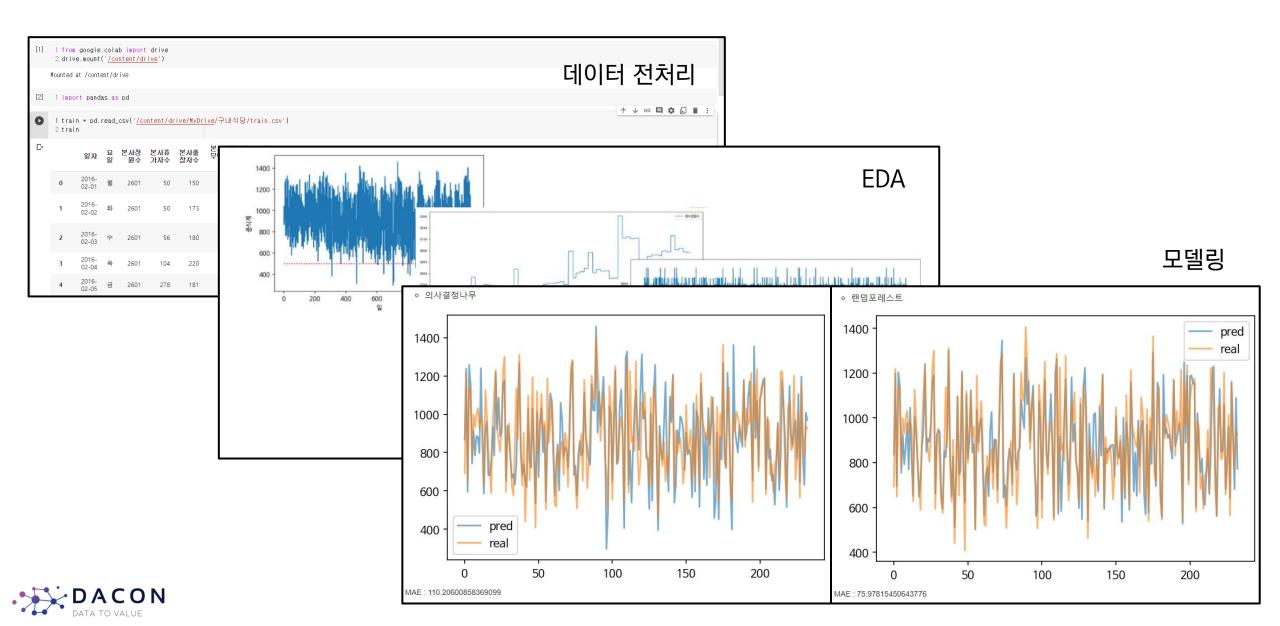


김세상

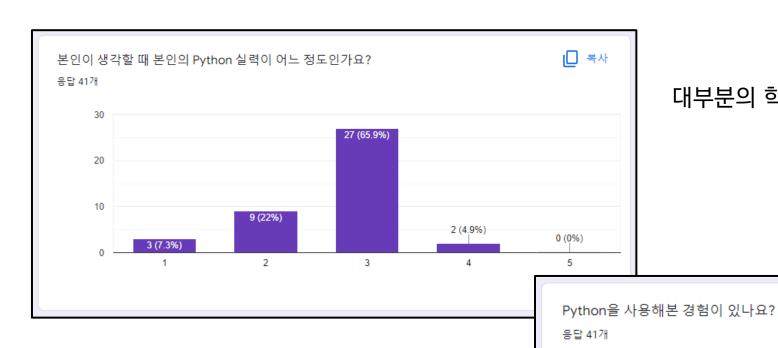
DACON / DS / MANAGER 기술 데이터 공학 전공 JOISS 빅데이터 경진대회 수상



### 우리는 두 달 동안...



### 여러분들은...



대부분의 학생들이 보통 수준의 Python 실력

1년 미만1년 이상

□ 복사

Python 사용 경험은 대부분 1년 이상

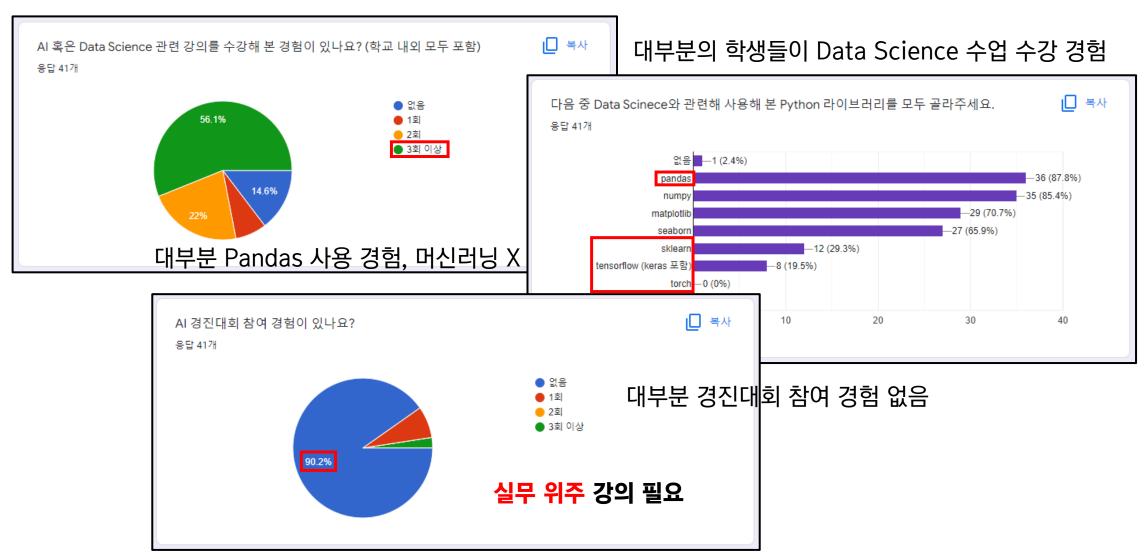
Python 활용 Project 경험 적음

85.4%

14.6%



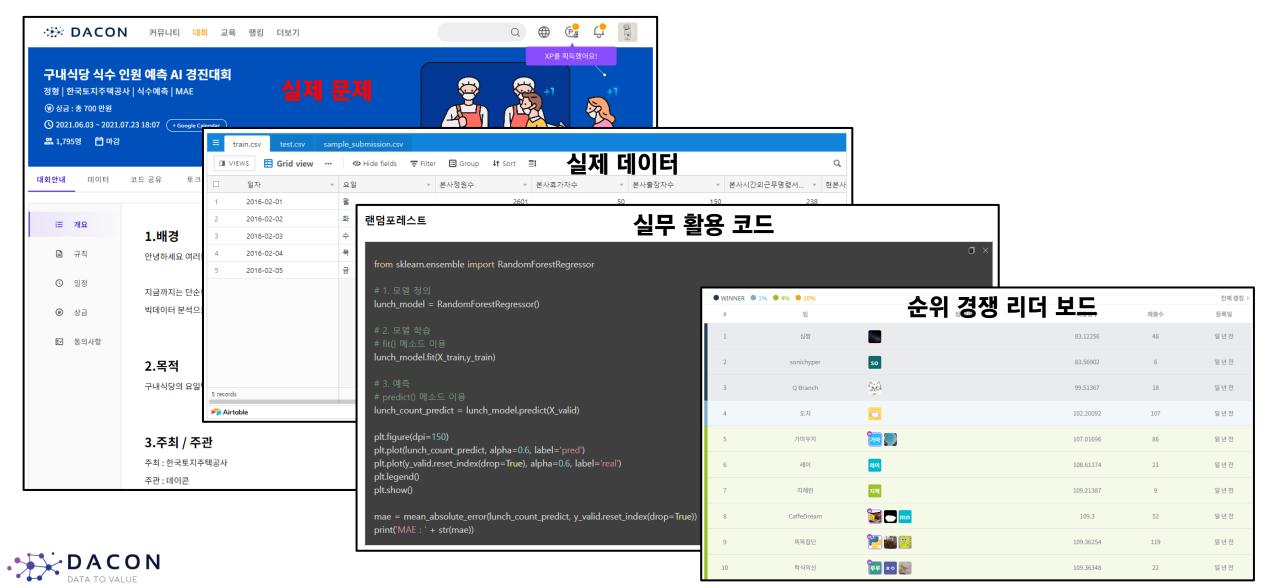
### 여러분들은...





#### 그래서 우리는...

#### **Project Based Learning**



### 근데 여러분들은...

#### 기초반

기초를 자세히 알려줬으면 좋겠다 열정반 기본부터 탄탄하게 다질 수 있게 해주시면 감사하겠습니다 자세한 설명을 부탁드립니다! 실습이 많았으면 좋겠습니당! 자세하고 쉽게 설명해주세요~ 구체적으로 잘 배우고 싶습니다. 많이 알려주시먄 김사하겠습니다 앞으로 잘 부탁드립니다:) 내가 이그래도 보통 이상의 수준은 하는 구나 라고 느끼고 싶습니다. 열심히 하겠습니다 너무 어렵지 않았으면 좋겠습니다! 배운 것을 잘 활용했으면 좋겠습니다 실무반 많이 배우고 싶습니다. 차근차근 많이 알려주시면 감사하겠습니다

수업만 듣다보니 실무에 관련된 내용을 잘 알지 못하는데 이번 기회에 실무에서는 어떤 역량을 주로 사용하는 지 알고싶습니다.

고인물

실력이 향상되는 계기가 되었으면 좋겠습니다!

실무경험을 채워줄 수 있는 좋은 강의가 되었으면 좋겠습니다:)

코딩 실력이 향상될수있기를 바라고 있습니다!

바로 공모전에 나갈 실력이 생겼으면 좋겠습니다!

교수님들의 홍보로 데이콘에 대해서 꾸준히 관심을 가지고 있었고, 교육 카테고리에서 #오늘의 파이썬이라는 이름으로 제공되는 학습 자료도 제게 데이터분석의 큰 흐름을 이해하는데 도움이 많이 되었습니다. 그러나 실제 경진대회 참여에 대해 막연한 두려움이 있었는데, 이번 프로그램을 계기로 데이콘에서 주관하는 체계적인 교육과 함께 경진대회 경험까지 해볼 수 있을거라는 생각에 매우 기대됩니다! 유익한 프로그램에 참여할 수 있게 되어 기쁩니다. 감사합니다:)

#### DACON

#### 그래서 강의는...

1회차

실습환경세팅, 파이썬 기초 문법, 파이썬 활용

2회차

판다스 기초 문법(map, groupby), 시각화(matplotlib)

3회차

식수 PBL 1 (시각화를 활용한 EDA)

4회차

식수 PBL 2 (EDA를 통한 인사이트를 바탕으로 모델 성능 개선) (대회 참여 1)

5회차

따름이 PBL 1 (시각화를 활용한 EDA)

6회차

따름이 PBL 2 (EDA를 통한 인사이트를 바탕으로 모델 성능 개선) (대회 참여 2)

7회차

대회1 대회2 솔루션 발표, 솔루션 코멘트

8회차

버스 PBL 1 (시각화를 활용한 EDA)

9회차

버스 PBL 2 (validation 셋으로 교차 검증, 파라미터 튜닝) (대회 참여 3)

10회차

병원 PBL 1 (시각화를 활용한 EDA)

11회차

병원 PBL 2 (앙상블 기법의 활용) (대회 참여 4)

12회차

대회3 대회4 솔루션 발표, 솔루션 코멘트, 시상식, 장학금 전달, 회식, 인턴



### 강의 내용















높은 점유율

프로그래밍 사이트 선정 상위 점유율 프로그래밍 언어 목록										
[IEEE Spectrum 2021 ]										
☞IEEE Spectrum에서 집계한 2021년 기준, 웹 분야 상위 10개 프로그래밍 언어										
1	Python	2	Java	3	Javascript	4	C#	5	Go	
6	HTML	7	PHP	8	Dart	9	Ruby	10	Rust	
☞IEEE Spectrum에서 집계한 2021년 기준, 모바일 분야 상위 10개 프로그래밍 언어										
1	Java	2	С	3	C++	4	C#	5	Swift	
6	Dart	7	Kotlin	8	Scala	9	Objective-C	10	Delphi	
	[ Stack Overflow 2021 ]									
	❷ Stack Overflow에서 조사한 2021년 기준 사용률 상위 25개 프로그래밍 언어									
1	JavaScript	2	HTML, CSS	3	Python	4	SQL	5	Java	
6	Node.js	7	TypeScript	8	C#	9	Bash	10	C++	
11	PHP	12	С	13	PowerShell	14	Go	15	Kotlin	
16	Rust	17	Ruby	18	Dart	19	어셈블리어	20	Swift	
21	R	22	VBA	23	MATLAB	24	Groovy	25	Objective-C	
	[ TIOBE 2022 ]									
	●TIOBE에서 선정한 2022년 2월 기준 검색이 점유율 상위 20개 프로그래밍 언어									
1	Python	2	С	3	Java	4	C++	5	C#	
6	Visual Basic .NET	7	JavaScript	8	PHP	9	어셈블리어	10	SQL	
11	Classic Visual Basic	12	R	13	Go	14	Fortran	15	Groovy	
16	Swift	17	Ruby	18	Perl	19	MATLAB	20	Delphi / Object Pascal	
	[ 21위 ~ 50위 펼치기 · 겁기 ]									
	[ PYPL 2022 ]									
	☞PYPL에서 선정한 2022년 3월 기준 검색어 점유율 상위 20개 프로그래밍 언어									
1	Python	2	Java	3	JavaScript	4	C#	5	C/C++	
6	PHP	7	R	8	Objective-C	9	TypeScript	10	Swift	
11	MATLAB	12	Kotlin	13	Go	14	Rust	15	Ruby	
16	VBA	17	Ada	18	Abap	19	Dart	20	Visual Basic	



대화형 인터프리터





```
Python 3.9.12 (main, Apr 4 2022, 05:22:27) [MSC v.1916 64 bit (AMD64)] :: Anaconda, Inc. on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> 1+2
3
>>> 'hi'*3
'hihihi'
>>> if 3 > 2 : print(True)
...
True
>>> __
```



#### 영어랑 비슷한 문법



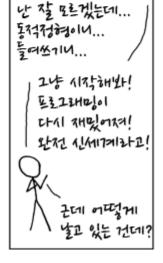
```
1 for i in range(0, 10):
2 | if i % 2 == 0:
3 | print('짝수')
4 | else:
5 | print('홀수')
5 | return x * factorial(x - 1)
```

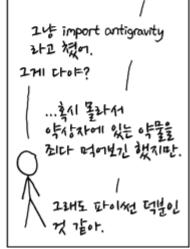


쉬운 난이도, 빠른 개발 속도









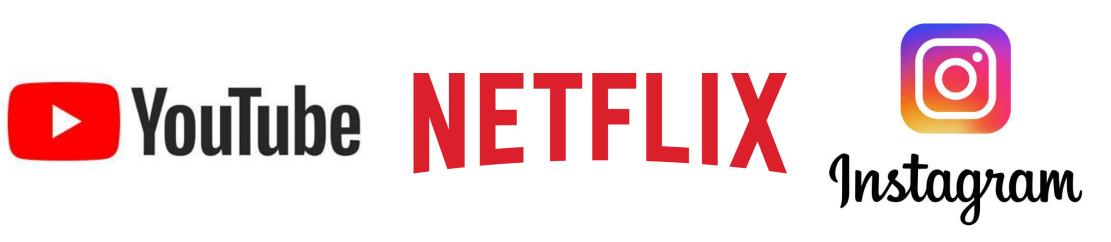




### 파이썬이란? 파이썬으로 만든 서비스







어떤 공통점?



파이썬으로 만든 서비스





어떤 공통점?



### **파이썬이란?** 우리는 파이썬으로…









O PyTorch

















### **실습 환경 세팅** 파이썬



```
>>> a = [[1, 2], [3, 4], [5, 6]]
>>> b = a.copy()
>>> id(a), id(b)
(4431577352, 4431577288)
>>> b[1]
[3, 4]
>>> b[1][1] = 0
>>> b
[[1, 2], [3, 0], [5, 6]]
>>> a
[[1, 2], [3, 0], [5, 6]]
>>> id(a[1]), id(b[1])
(4431577736, 4431577736)
>>> import copy
>>> a = [[1, 2], [3, 4], [5, 6]]
>>> b = copy.copy(a)
>>> id(a), id(b)
(4431556232, 4431717128)
>>> id(a[1]), id(b[1])
(4431556104, 4431556104)
>>> b = copy.deepcopy(a)
>>> id(a), id(b)
(4431556232, 4431577288)
>>> id(a[1]), id(b[1])
(4431556104, 4431577608)
```

```
.Join(comments)).split('|')
if tar_word in list_word and seed_word_p in list_word:
    if count_word <= NEAR: # if number of word in each comment lass the</pre>
         positive_tar = []
         positive_sw = []
         for l_index, l_word in enumerate(list_word):
             if tar_word - 1_word:
                  positive_tar.append(l_index)
              elif seed word p - 1 word:
                  positive_sw.append(l_index)
         breaker = False
         for index_tar in positive_tar: # calculate distance between an
              for index_sw in positive_sw:
                  if ((index_tar - index_sw) <- NEAR) and ((index_tar
                      count_p += 1
                      breaker = True
```



ANACONDA(Jupyter) Colab











### 실습 환경 세팅 ANACONDA(Jupyter) 설정 - 계정

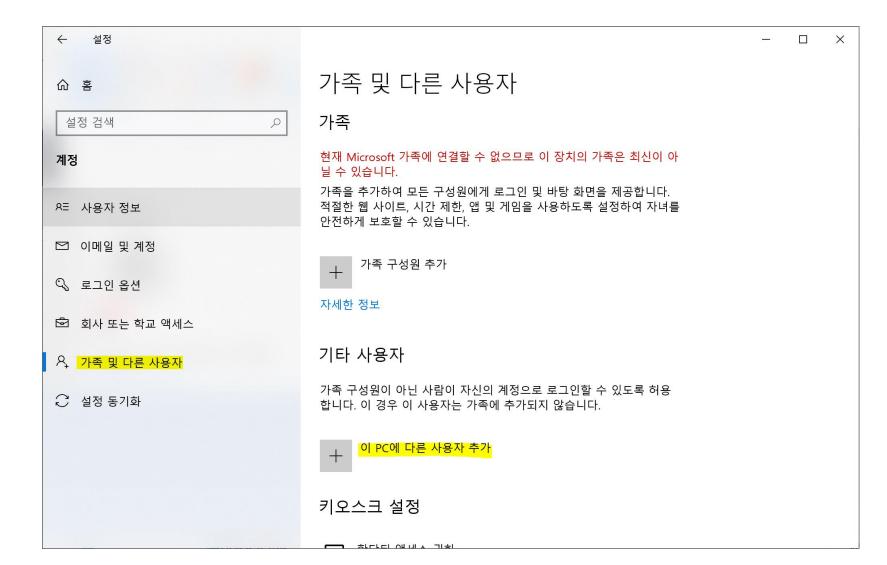


설정						_	×
		설정 검색		ρ			
旦	<b>시스템</b> 디스플레이, 소리, 알림, 전원	E	<b>장치</b> Bluetooth, 프린터, 마우스		<b>전화</b> Android, iPhone 연결		
	<b>네트워크 및 인터넷</b> Wi-Fi, 비행기 모드, VPN	<b>£</b>	<b>개인 설정</b> 배경, 잠금 화면, 색	≣	<b>앱</b> 설치 제거, 기본값, 옵션 기능		
8	<b>계정</b> 내 계정, 메일, 동기화, 회사, 가 족	心 A字	<b>시간 및 언어</b> 음성, 지역, 날짜	8	<b>게임</b> Xbox Game Bar, 캡처, 게임 모 드		
G.	<b>접근성</b> 내레이터, 돋보기, 고대비	٥	<b>검색</b> 내 파일 찾기, 사용 권한		<b>개인 정보</b> 위치, 카메라, 마이크		
	업데이트 및 보안 Windows 업데이트, 복구, 백업						



ANACONDA(Jupyter) 계정 - 가족 및 다른 사용자

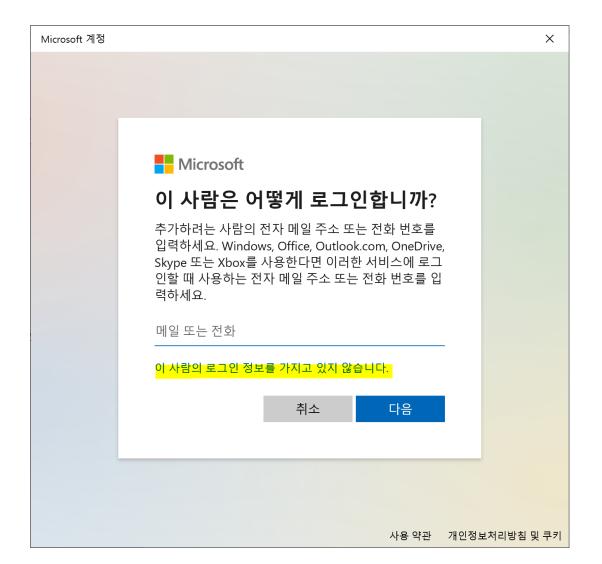






#### 실습 환경 세팅 ANACONDA(Jupyter) 다른 사용자 추가

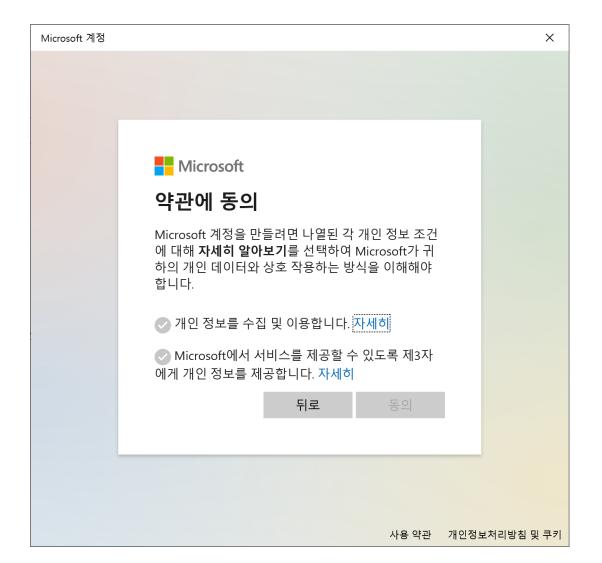






#### 실습 환경 세팅 ANACONDA(Jupyter) 약관에 동의

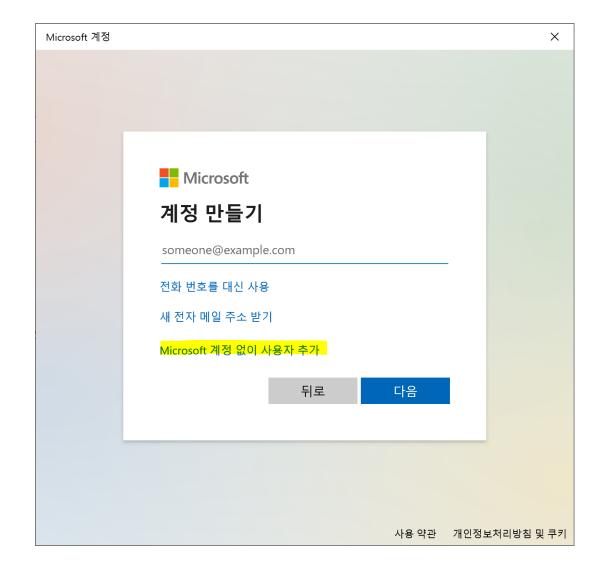






### 실습 환경 세팅 ANACONDA(Jupyter) 계정 없이 사용자 추가

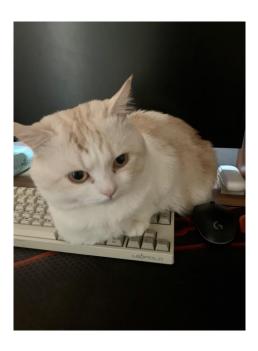






ANACONDA(Jupyter) 계정 없이 사용자 추가







Microsoft 계정 ×
11 10 1 LE 1
이 계정이 어린이나 청소년을 위한 계정인 경우 <b>뒤로</b> 선택하고 Microsoft 계정 만드는 것이 좋습니다. 젊은 가족 구성원이 Microsoft 계정 로그인하면 연령에 중점을 두는 개인 된 보호가 적용됩니다.
암호를 설정할 때는 본인은 기억하기 쉬우면서 다른 사람들이 추측하기 어려운 암호를 용하십시오. 계정 이름은 반드시 영어로!!
이 PC를 누가 사용하나요?
Dacon
보안 암호를 만듭니다.
••••
••••
비밀번호를 잊어버린 경우
첫 번째 애완 동물의 이름은 무엇이었습니까? ~
돼지
태어난 도시의 이름은 무엇입니까?
부산
어린 시절 별명은 무엇입니까?
돼지

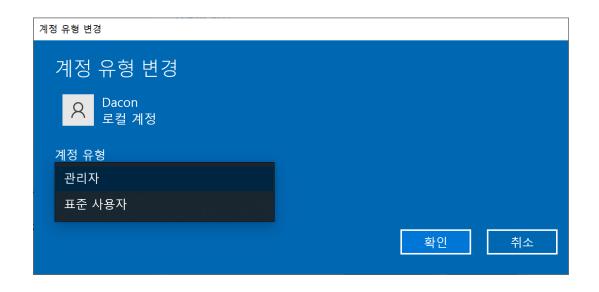
ANACONDA(Jupyter) 가족 및 다른 사용자 - 계정 유형 변경







ANACONDA(Jupyter) 가족 및 다른 사용자 - 계정 유형 변경









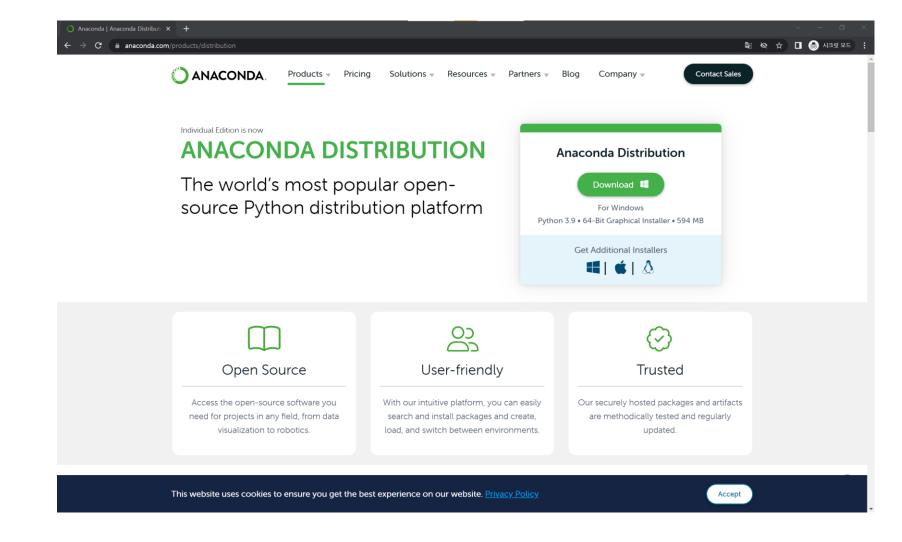
ANACONDA(Jupyter) 시작 - 계정 변경 - 로그인



۶ <sub>4</sub>	계정 설정 변경
	잠금
<u></u>	로그아웃
8	Dacon
8	
	문서
	사진
£(j);	설정
O	전원

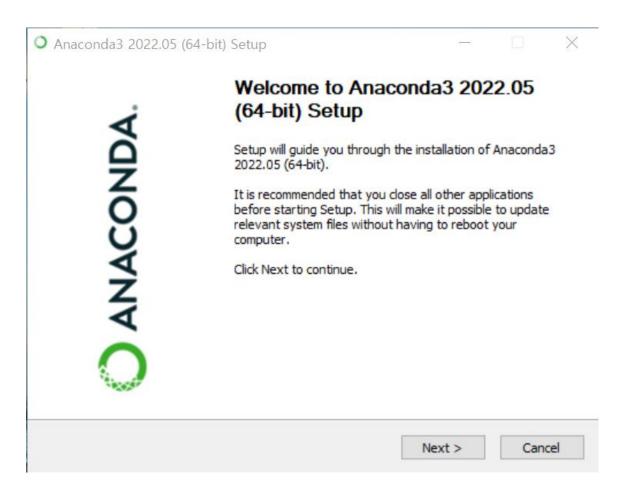








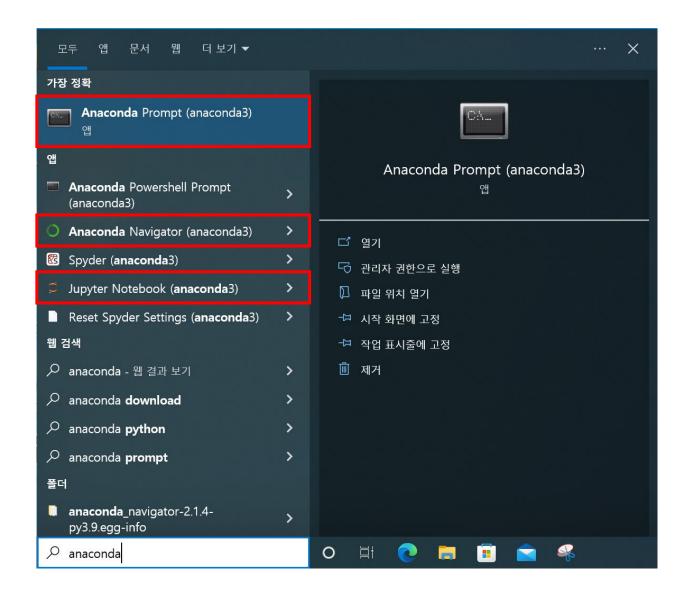




	License Agree	ement				
) ANACONDA.	Please review the license terms before installing Anaconda3 2022.05 (64-bit).					
Press Page Down to see th	e rest of the agre	ement.				
	at Assessed Di				^	
End User License Agreeme	ent - Anaconda Dis	stribution ========		===		
Copyright 2015-2022, Ana	aconda, Inc.					
All rights reserved under t	he 3-dause BSD L	icense:				
This End User License Agrand Anaconda, Inc. ("Anawas formerly known as An	conda") and gove	rns your use of A			~	
If you accept the terms of agreement to install Anaco			tinue. You must	accept the		
aconda, Inc. ————						
		< Back	I Agree	Cano		

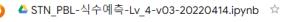








### 실습 환경 세팅 Google Colab



a. 이슈 도출 및 문제 정의

a. Files - 파일 불러오기 b. 데이터 살펴보기

a. 탐색적 자료 분석 - EDA

본사휴가자수

본사출장자수

중식계 석식계

본사시간외근무명령서승인건수

현본사소속재택근무자수

가설 검증을 위한 시각화

을 받을 것이다.

어질 것이다.

Step 4. 모델링

b. 데이터 처리 Process

a.1. 중식계 모델 정의 a.2. 중식계 모델 학습 / 예측

b.1. 석식계 모델 정의

b.2. 석식계 모델 학습 / 예측

1.식수인원은 본사정원수에 큰 영향

scatter plot(산점도)이란?

휴가자, 출장자, 재택근무자는 구내

식당에서 식사를 하지 않기 때문에 해당 수치가 높다면 식수 인원은 적

b. 가설 설정 c. 분석 실험 설계 Step 2. 데이터 준비

Step 3. 데이터 가공

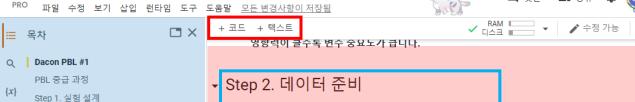












▼ a. Files - 파일 불러오기

데이터를 불러오기 위해서는 pandas 패키지를 불러와야겠죠?

[] 1 import pandas as pd

다음으로 지난 시간과 마찬가지로 대회 데이터를 불러와 보겠습니다. 해당 대회의 파일은 총 3가지 이며 모두 csv파일입니다.

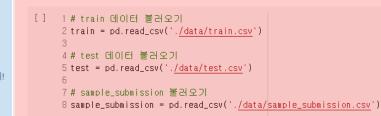
- train.csv : 모델을 학습하는 데 사용할 데이터
- test.csv: 모델을 통해 최종 제출할 식수인원을 예측하는 데 사용할 데이터
- sample\_submission.csv: test.csv 파일과 모델을 이용해 식수인원을 예측한 뒤 해당 값을 기 록하여 제출하는 제출용 파일

대회에 참가하는 참가자라면 아래와 같은 프로세스로 데이터 분석을 진행하게 됩니다.

- 1. train 데이터로 모델 학습
- 2. 교차 검증을 통해 모델의 성능을 확인하여 모델이 과적합 되었는지 확인하며 모델 최적화
- 3. 2번에서 최적화된 모델을 바탕으로 test 데이터 예측
- 4. 3번의 예측 값으로 sample\_submission의 값을 변경하고 변경된 sample\_submission 파일을 데이콘 홈페이지에 제출

◘2번 과정은 현재 과정에서는 다소 어려운 내용이기 때문에 다음 과정에서 알아보도록 하겠습니다 중급 과정에서는 1, 3, 4번 과정을 다루게 됩니다.

pandas의 read\_csv() 함수를 이용해 분석에 사용할 데이터를 불러옵니다.





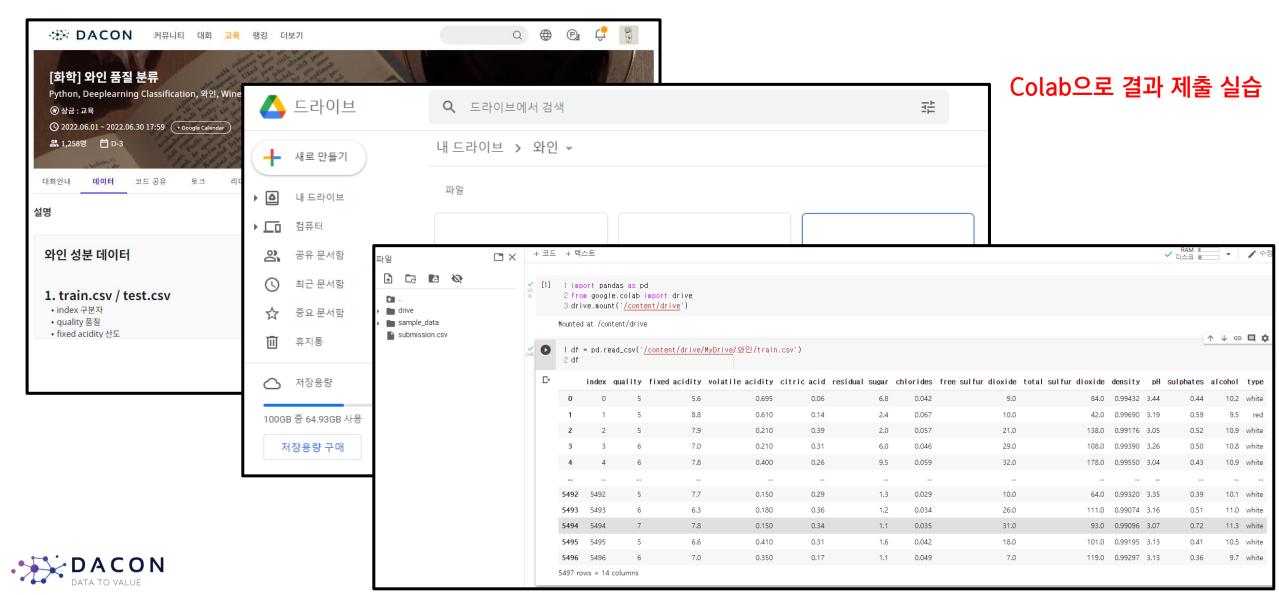




### Google Colab

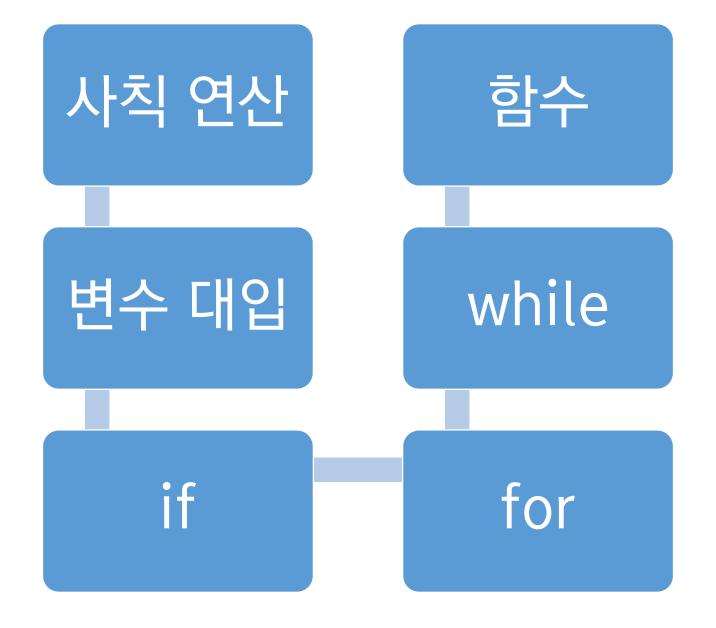
대회 참여 - 결과 제출













### **파이썬** 자료형



#### 숫자

• 정수 : 123, -456,

• 실수 : 123.45

• 8진수, 16진수

#### 문자열

- 'python'
- '123'

#### 리스트

- [1,2,3,4,5]
- ['정성문', '김세 상']

#### 튜플

- (1,2,3,4,5)
- ('정성문', '김세 상')
- 변경 불가

#### 딕셔너리

- {1: '정성문', 2:' 김세상}
- Key, value

#### 집합

- {1,2,3,4,5}
- 중복 X
- 순서 X

#### 불

- True
- False



### **파이썬** 자료형 연산



#### 숫자

• 사칙연산(+ - \* /), 제곱연산(\*\*), 나머지 연산(%), 몫 연산(//)

#### 문자열

• 더하기, 곱하기, 길이반환(len), 인덱싱([4]), 슬라이싱([2:4])

#### 리스트

• 더하기, 곱하기, 인덱싱, 슬라이싱

#### 튜플

• 리스트와 같음

#### 딕셔너리

Keys, values, items

#### 집합

• 교집합(&), 합집합(I), 차집합(-)

#### 불

• ==, \langle, \rangle, \langle =, \rangle =



#### 자료형 연습문제 1



```
주어진 딕셔너리에 다음과 같은 지시를 따르시오
```

```
[] 1 students = {'이지은':100,
2 | '강동원':95,
3 | '송강호':90,
4 | '배두나':85,
5 | '이주영':95}
```

#### 문제1

- 1. 딕셔너리에서 점수를 리스트로 출력하시오.
- 2. 집합을 이용하여 중복값을 제거해 보시오.

```
정답 1번: [100, 95, 90, 85, 95]
2번: {90, 100, 85, 95}
```



#### 자료형 연습문제 2



```
주어진 딕셔너리에 다음과 같은 지시를 따르시오
```

```
[] 1 students = {'이지은':100,
2 | '감동원':95,
3 ' '송강호':90,
4 '배두나':85,
5 | '이주영':95}
```

#### 문제2

- 1. 이지은 학생의 점수를 문자열 'A+'으로 바꾸시오.
- 2. items() 함수를 이용하여 딕셔너리를 확인하시오

정답

dict\_items([('이지은', 'A+'), ('강동원', 95), ('송강호', 90), ('배두나', 85), ('이주영', 95)])



### **파이썬** 제어문



#### lf문

- if
- elif
- else

#### While문

break

#### for문

continue

#### 리스트

#### 컴프리헨션

• 제어문을 중첩하여 리스트 생성

#### 함수

• 데이터를 자유자재로 조작하기 위한 커스텀 함수

#### 내장함수

- int
- len
- range
- list
- map ...



### **파이썬** 제어문 연습문제 1



문제1. fruit\_list를 생성하여 다음과 같은 지시를 따르시오.

1) 반복문을 사용하여 과일 이름을 key로 과일의 순서를 value로 갖는 fruit\_dict를 생성하시오.

[2] 1 fruit\_list = ['오렌지', '사과', '딸기', '배', '수박']

정답

7 fruit\_dict

{'딸기': 2, '배': 3, '사과': 1, '수박': 4, '오렌지': 0}



#### 제어문 연습문제 2



문제2. fruit\_dict를 사용하여 다음의 지시를 따라 fruit\_list를 수정하시오.

[0, 1, 2, 3, 4]

1) map 함수를 사용하여 fruit\_list의 과일이름을 fruit\_dict의 숫자로 바꾸시오.





#### 제어문 연습문제 3



문제2. fruit\_dict를 사용하여 다음의 지시를 따라 fruit\_list를 수정하시오.

1) map 함수를 사용하여 fruit\_list의 과일이름을 fruit\_dict의 숫자로 바꾸시오.





#### 실습문제 1



#### 1번 문제

- 홀짝 임의의 자연수 x가 주어질 때, x가 홀수인지 짝수인지 판별하는 함수를 작성하세요
- 입출력 예

X	result				
10	'짝수'				
11	'홀수'				





#### 실습문제 2

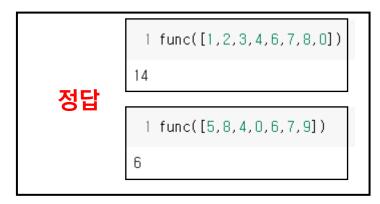


#### 2번 문제

• 없는 숫자 더하기

0부터 9까지의 숫자 중 일부가 들어있는 정수 배열 numbers가 매개변수로 주어집니다. numbers에서 찾을 수 없는 0부터 9까지의 숫자를 모두 찾아 더한 수를 return 하도록 func 함수를 완성해주세요.

• 입출력 예 numbers result [1,2,3,4,6,7,8,0] 14 [5,8,4,0,6,7,9] 6



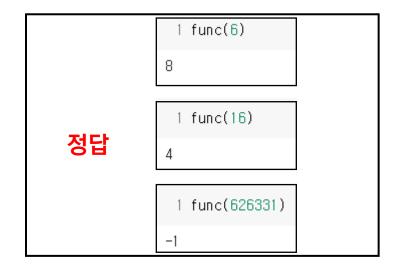


### **파이썬** 실습문제 3



#### 3번 문제

- 콜라츠 추측
- 1-1). 입력된 수가 짝수라면 2로 나눕니다.
  - 1-2). 입력된 수가 홀수라면 3을 곱하고 1을 더합니다.
  - 2). 결과로 나온 수에 같은 작업을 1이 될 때까지 반복합니다.
- 단, 주어진 수가 1인 경우에는 0을, 작업을 500번 반복할 때까지 1이 되지 않는다면 -1을 반환해 주세요.
- 입출력 예 numbers result 6 8 16 4 626331 -1





#### 실습문제 4



#### 4번 문제

- 하샤드 수
- 양의 정수 x가 하샤드 수이려면 x의 자릿수의 합으로 x가 나누어져야 합니다. 예를 들어 18의 자릿수 합은 1+8=9이고, 18은 9로 나누어 떨어지므로 18은 하샤드 수입니다. 자연수 x를 입력받아 x가 하샤드 수인지 아닌지 검사하는 함수, solution을 완성해주세요.
- 입출력 예 numbers result
  10 True
  12 True
  11 False
  13 False

