orange 활용 데이터 분석 및 머신 러닝

목차

1차시

데이터분석과 orange !

데이터 분석이란?
데이터 시각화란?

데이터와 정보 (DIKW 피라미드)

지식을 활용하는 창의적인 아이디어

정보들 간의 관계를 통해 얻은 가치 있는 정보

유의미하게 가공된 2차 데이터

가공하기 전의 순수한 상태의 자료들

데이터와 정보 (DIKW 피라미드)

눈오는 날 10% 할인 판촉으로 매출 더욱 높임
다른 계절도 날씨에 따른 매출 영향을 분석하여 수요 발견

눈오는 날 매출이 증가함.
배달이 되지 않아 근처 회사의 take-out 손님이 몰림.

일간 매출, 주간 매출, 월간 매출,
일 평균 매출

맥도널드 지점의 매출 데이터

데이터와 정보 (DIKW 피라미드)

A마트의 다른 상품들도 B마트보다 쌀 것이라고 예측

연필을 살 때는 A마트로 가는 것이 유리

A마트의 연필가격이 더 저렴함

A마트 펜 500원, 연필 200원, 라면 3000원..
B마트 연필 300원, 라면 2700원…

데이터 분석이 필요한 이유

1) 데이터는 모든 현상과 가설의 '근거'가 되는 정보를 담고 있음

2) 직접 경험하지 않아도 데이터를 통해 '경험'을 얻을 수 있음

3) 데이터분석을 통해 데이터에서 의미를 찾고 미래를 예측하는 등 새로운 가치를 창출할 수 있음.

aimap marker May 21, 2019 · 9 min read

https://medium.com/@aimap.marke

비즈니스에 대한 정보가 매 순간 데이터로 기록

↓

데이터의 종류, 양, 접근성 ↑

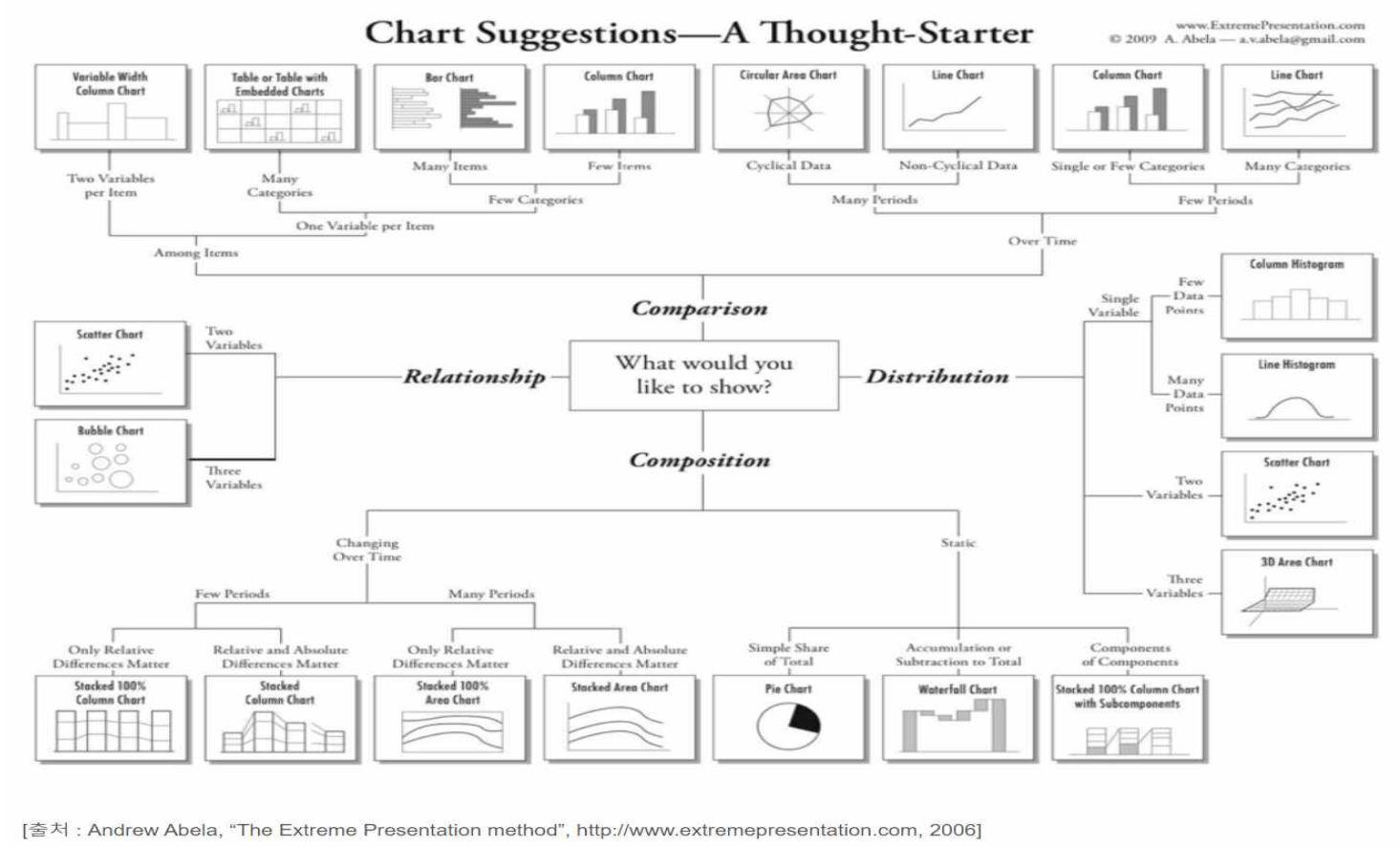
↓

데이터를 이해하고 활용할 수 있는 능력

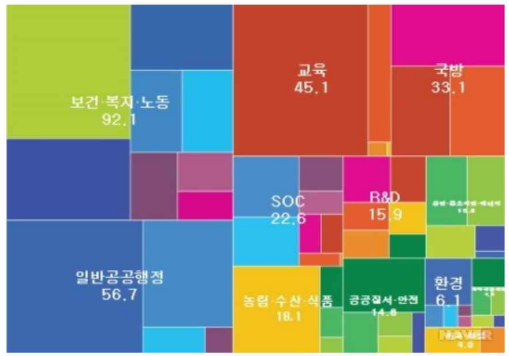
↓

데이터 분석을 통한 가치창출

데이터 시각화(Data Visualization)



[출처 : Andrew Abela, “The Extreme Presentation method”, http://www.extremepresentation.com, 2006]



인포그래픽으로 표현한 2012년 예산안 (단위 : 조원)

데이터 분석 프로세스

- 데이터 분석의 목적은 문제 해결

데이터 분석 프로세스

- 문제 해결에 필요한 데이터의 속성을 정의.
  데이터의 구체적인 정보 항목으로
  더 이상 분리될 수 없는 최소의 데이터 보관 단위

데이터 분석 프로세스

문제 정의

데이터 분석의 목적은 문제해결

데이터 정의

필요한 데이터의 속성은?

데이터 수집

설문, 공공데이터(2차원 표 형태) 등

데이터 전 처리

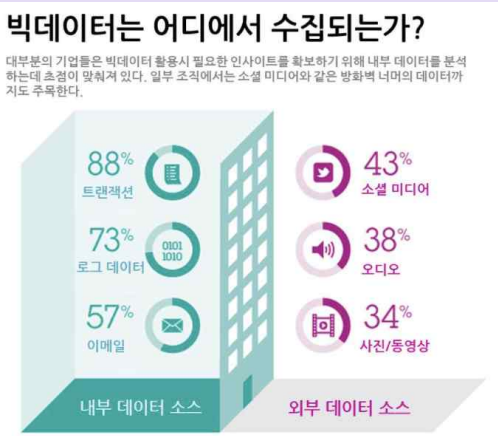
이상치, 결측치 처리 데이터 형태 변환

데이터 시각화

데이터 이해

결과해석

결론 도출, 문제해결



제공되는 데이터, 직접 수집, 웹 크롤링/웹스크래핑

IBM

공공데이터 활용지표 :
http://www.index.go.kr/potal/main/EachDtlPageDetail.do?idx_cd=2844

CSV	XLSX	JSON	XML	API
데이터를 콤마 (,)로 구분	엑셀 포맷 (행, 열)	Key:Value 구조	마크업	개발자를 위한 인터페이스

날짜,지점,평균,최저,최고
2021-07-24,108,31.7,26.9,36.5
2021-07-25,108,31.5,27.2,35.9
2021-07-26,108,31.2,27.4,35.4
2021-07-27,108,31.1,27.8,35.7
2021-07-28,108,30.4,27.1,34.7
2021-07-29,108,29.6,27,33.3
2021-07-30,108,30.5,25.8,35.4
2021-07-31,108,29.8,26.8,34.3
2021-08-01,108,27.1,25.1,28.8
2021-08-02,108,26.5,25,28.6

A	B	C
날짜	지점	평균기온(℃) 최
2021-07-24	108	31.7
2021-07-25	108	31.5
2021-07-26	108	31.2
2021-07-27	108	31.1
2021-07-28	108	30.4
2021-07-29	108	29.6
2021-07-30	108	30.5
2021-07-31	108	29.8
2021-08-01	108	27.1
2021-08-02	108	26.5
2021-08-03	108	28


```
{
  "menu":"회원관리",
  "member":[
            {
                "id":"dragon",
                "name":"홍길동",
                "address":"서울시",
                "job":"학생"
            },
            {
                "id":"dragun",
                "name":"홍길서",
                "address":"제주시",
                "job":"교사"
            }
    ]
}
```



```
<?xml version="1.0" encoding="euc-kr" ?>
- <거래내역서>
 - <내역>
    <물품번호>AA1</물품번호>
    <물품명>pen</물품명>
    <단가 unit="dollor">1.0</단가>
    <수량>30</수량>
    <총액 unit="dollor">30</총액>
    <거래연도>2001</거래연도>
    <거래월>01</거래월>
    <거래일>20</거래일>
  </내역>
 + <내역>
</거래내역서>
```


https://www.itworld.co.kr/news/78701

데이터 분석 프로세스

문제 정의

데이터 분석의 목적은 문제해결

데이터 정의

필요한 데이터의 속성은?

데이터 수집

설문, 공공데이터(2차원 표 형태) 등

데이터 전 처리

이상치, 결측치 처리 데이터 형태 변환

데이터 시각화

데이터 이해

결과해석

결론 도출, 문제해결

1) 해결하려는 문제에 필요한 속성만을 데이터셋에서 추출
2) 결과에 영향을 미치지 않도록
   결측값/이상값이 포함된 데이터는 삭제 하거나 대표할 수 있는 값으로 채운다(대표값, 평균값, 최빈값, 중앙값)

Garbage in garbage out!!

	A	B	C	D	E
1	날짜	지점	평균기온(°	최저기온(°	최고기온(°
2	2021-07-24	108	31.7	26.9	36.5
3	2021-07-25	108	31.5	27.2	35.9
4	2021-07-26	108	31.2	27.4	35.4
5	2021-07-27	108	31.1	27.8	35.7
6	2021-07-28	108	30.4	27.1	34.7

데이터 분석 프로세스

문제 정의

데이터 분석의 목적은 문제해결

데이터 정의

필요한 데이터의 속성은?

데이터 수집

설문, 공공데이터(2차원 표 형태) 등

데이터 전 처리

이상치, 결측치 처리 데이터 형태 변환

데이터 시각화

데이터 이해

결과해석

결론 도출, 문제해결

https://m.mt.co.kr/renew/view.html?no=2018080613120727477&ca=#_enliple

데이터 분석 프로세스

문제 정의

데이터 분석의 목적은 문제해결

데이터 정의

필요한 데이터의 속성은?

데이터 수집

설문, 공공데이터(2차원 표 형태) 등

데이터 전 처리

이상치, 결측치 처리 데이터 형태 변환

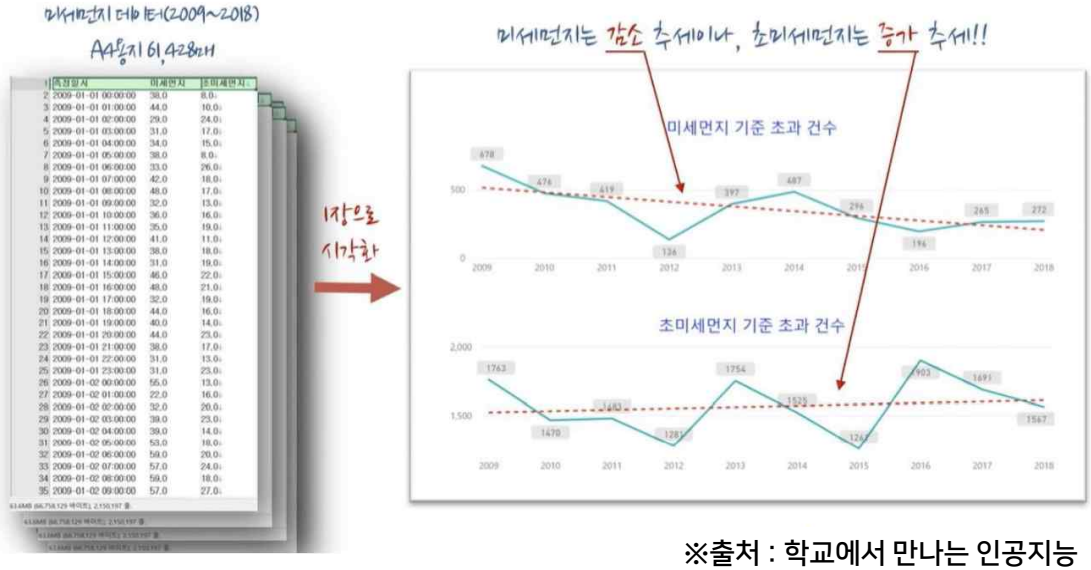
데이터 시각화

데이터 이해

결과해석

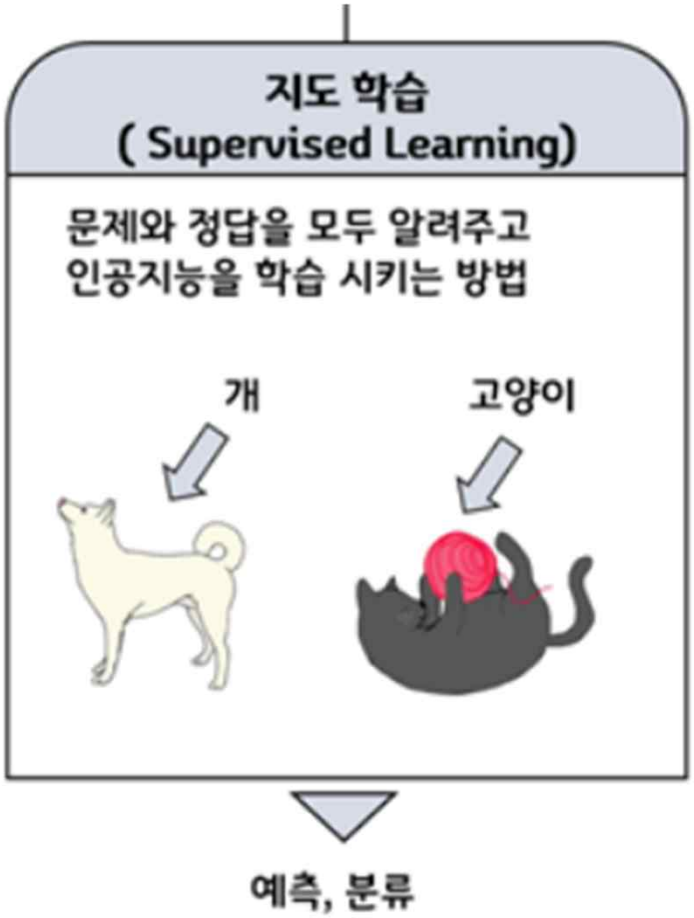
결론 도출, 문제해결

1) 많은 데이터를 한눈에 파악할 수 있다.
2) 데이터의 변화, 데이터 사이의 관계 등을 쉽게 볼 수 있어 데이터를 깊이 있게 이해할 수 있다.
3) 수치로만 파악하기 힘든 패턴이나 새로운 정보를 발견할 수 있다.
4) 인공지능이 학습할 데이터를 점검할 수 있다.
5) 인공지능이 바르게 일을 처리했는지 확인할 수 있다.



※출처 : 학교에서 만나는 인공지능

분석을 위한 데이터의 성격 정의하기



Target
Class
Label
종속변수
목푯값
정답
Category

VS.

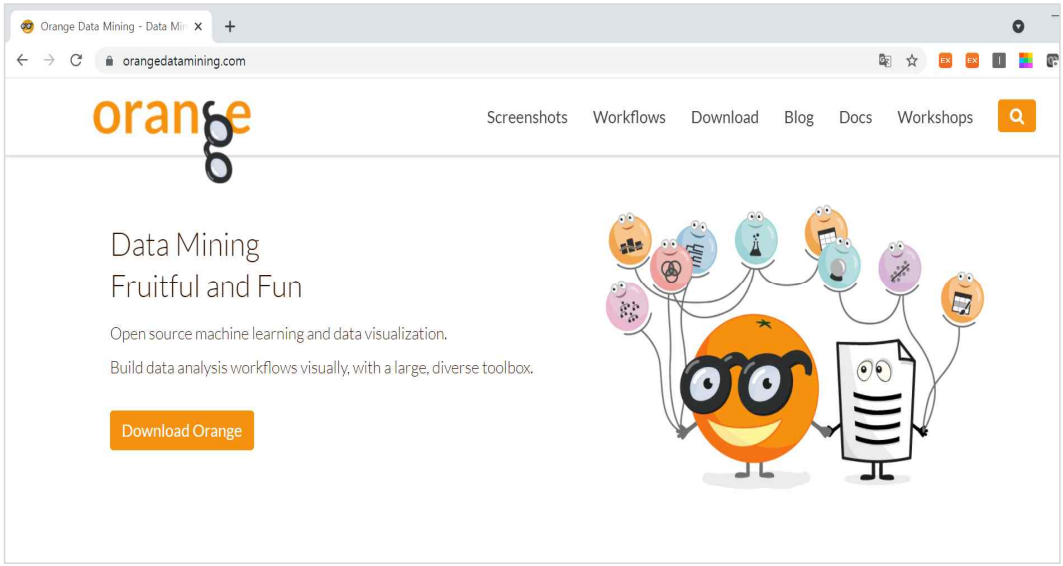
Feature
Attributes
독립변수
문제
속성

설치하기

Orange3 란?

- 오픈소스 데이터 시각화 및 머신러닝을 이용한 데이터 분석 도구
- 1997년 슬로베니아 루블라냐 대학에서 개발, 2015년 Orange 3.0 배포
- 데이터 분석을 위한 컴포넌트 기반 비주얼 프로그래밍 소프트웨어
- 데이터 분석 워크플로우를 시각적으로 구축

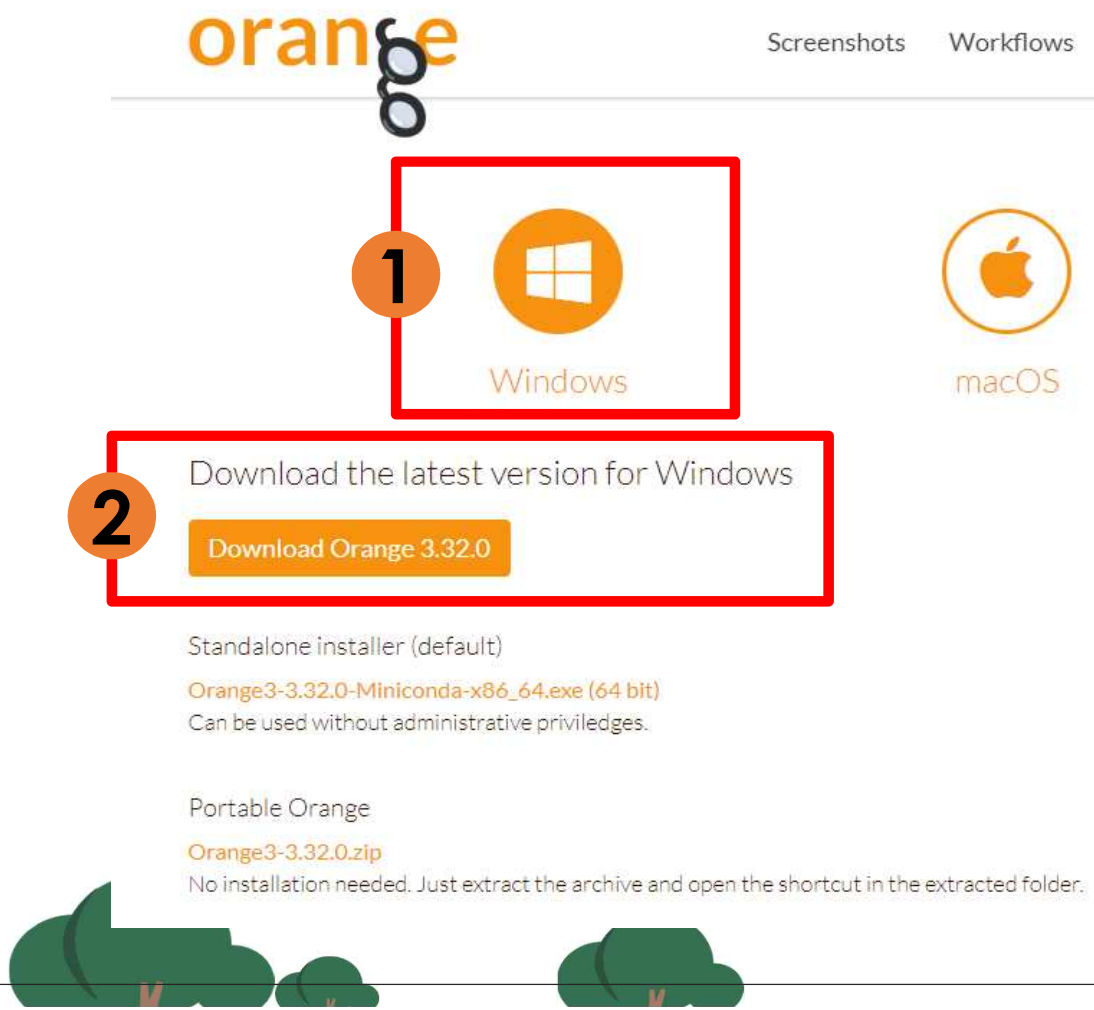
코딩과 수학 없이 드래그 앤 드롭으로
머신 러닝을 이용한 데이터 분석 가능



https://orangedatamining.com/

Orange 설치

https://orangedatamining.com/



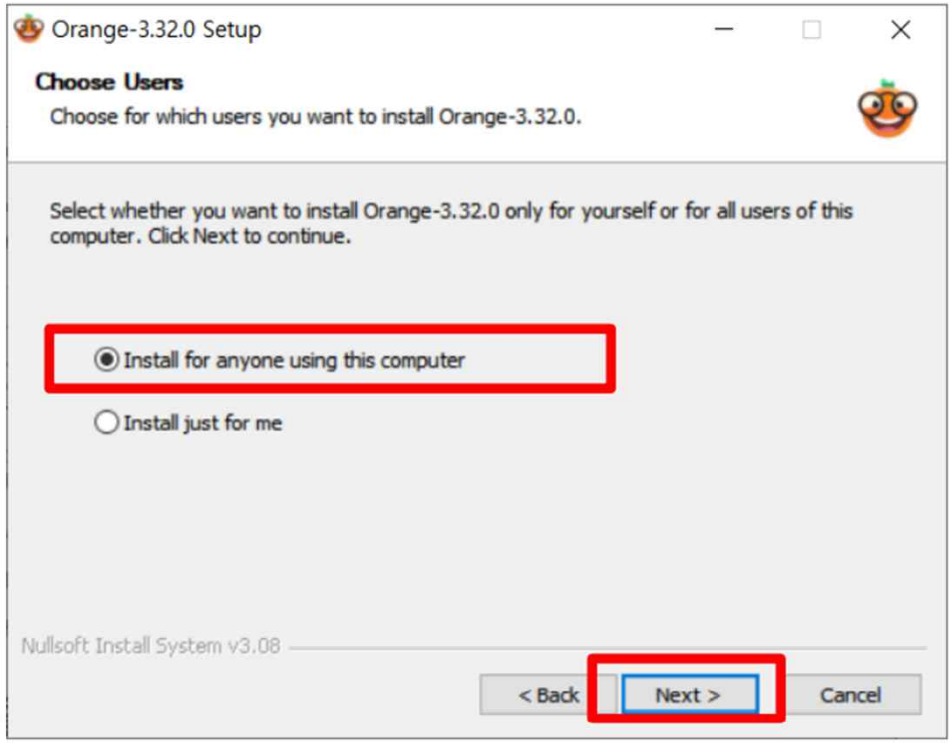
가장 최신버전으로 설치하세요^^
Orange3-3.33.0-Miniconda-x86_64.exe

설치파일은 533MB 정도이고, 설치를 모두 하고 나면 1GB가 훨씬 넘습니다.

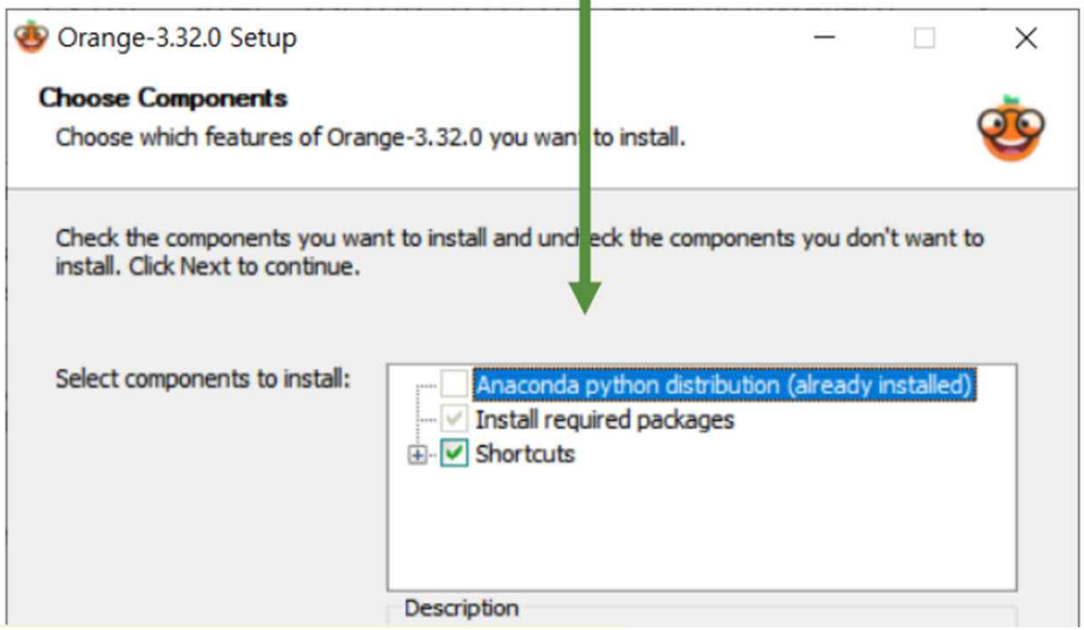
파일이 다운로드 되면 실행시켜 관리자권한으로 설치 시작합니다

Orange 설치

https://orangedatamining.com/



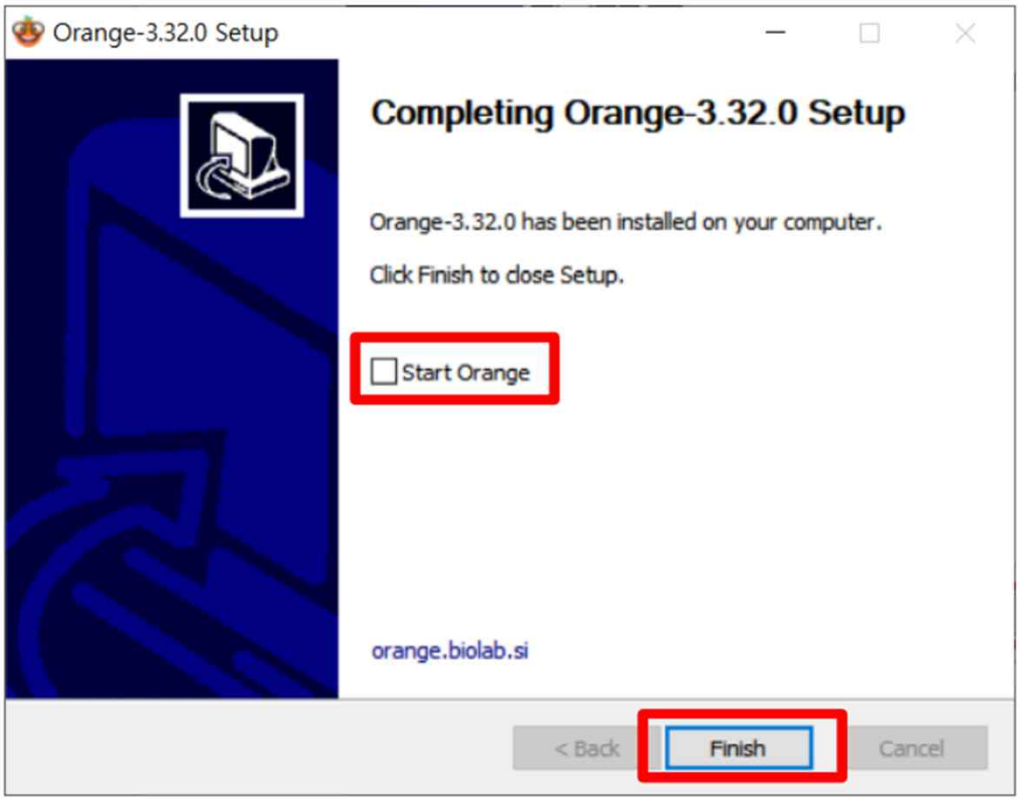
Anaconda가 설치되어 있지 않은 경우
(저는 이미 설치되어 있습니다. already installed)
Miniconda를 설치하는 과정이 더 있을 수 있습니다.
정상적인 과정이니 모두 설치하세요.



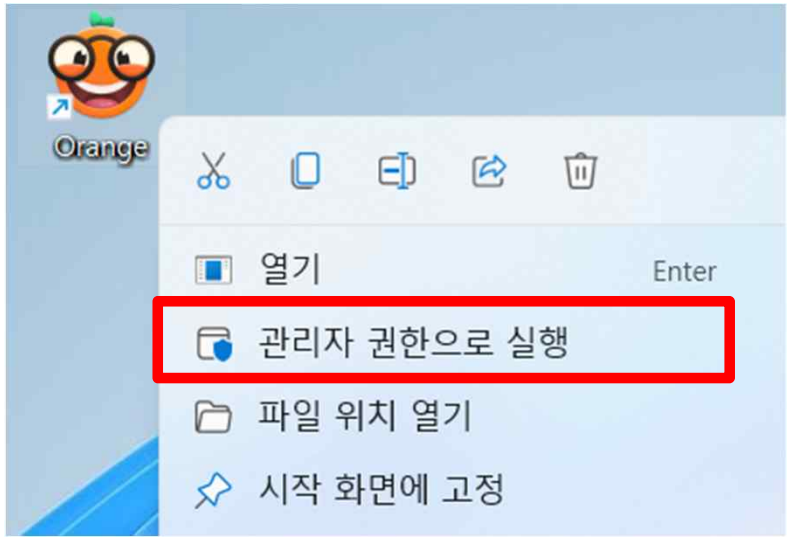
Orange 설치

https://orangedatamining.com/

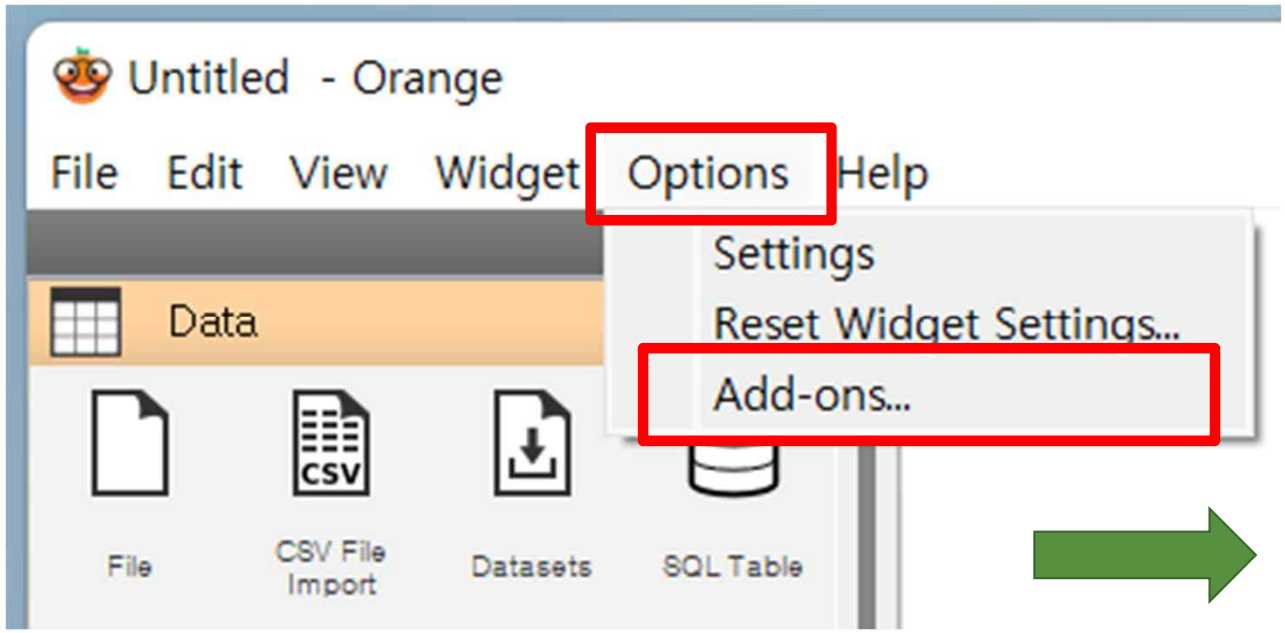
설치에 시간이 오래 걸립니다……

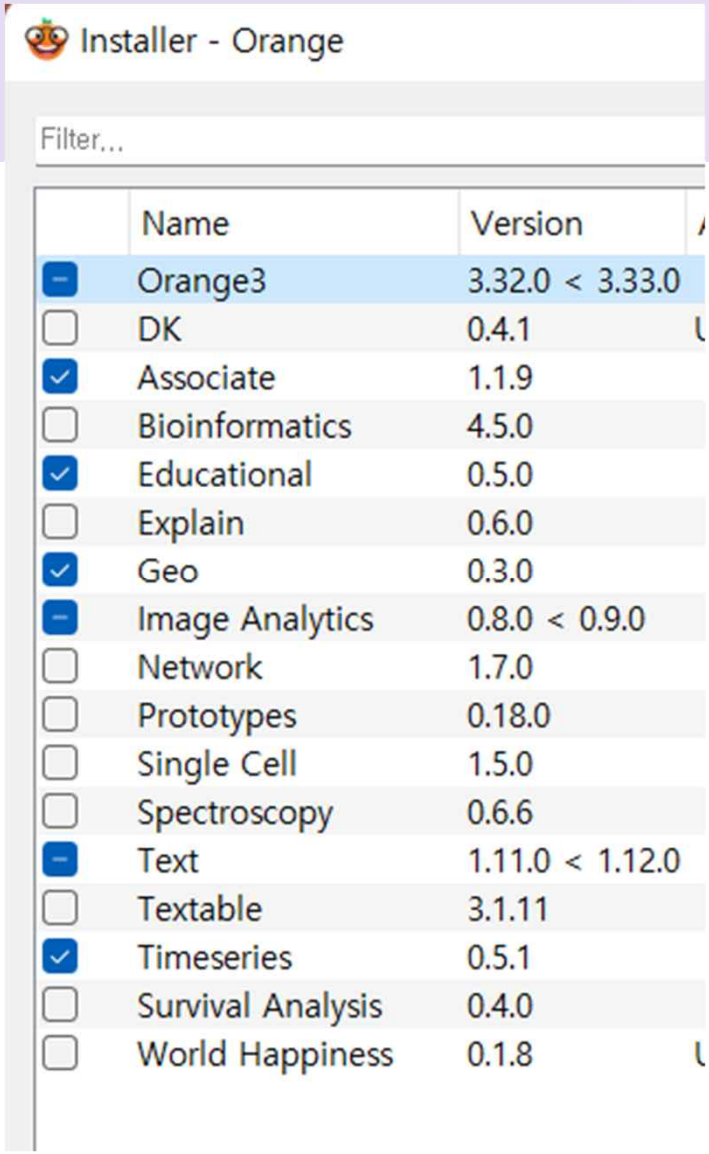


설치를 마치면 Add-on 설치를 위해 바탕화면에서 관리자 권한으로 오렌지를 실행합니다.



Orange3 Add-on 설치

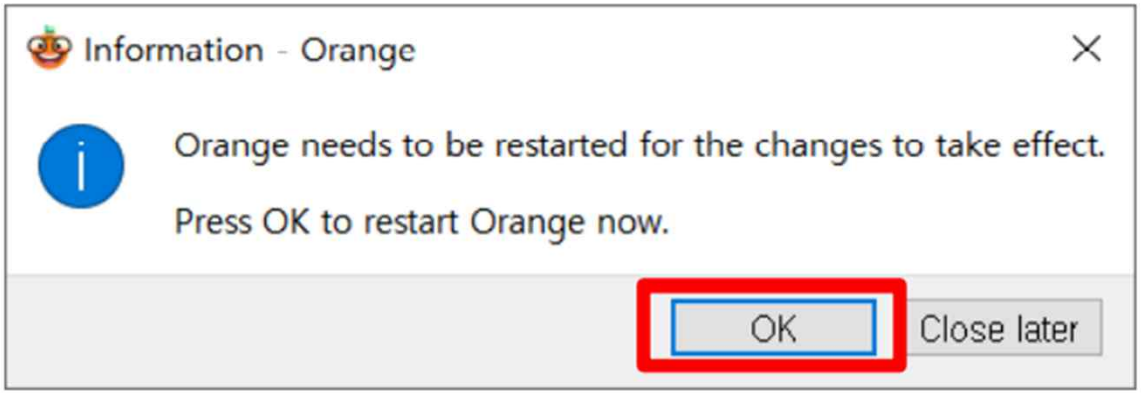


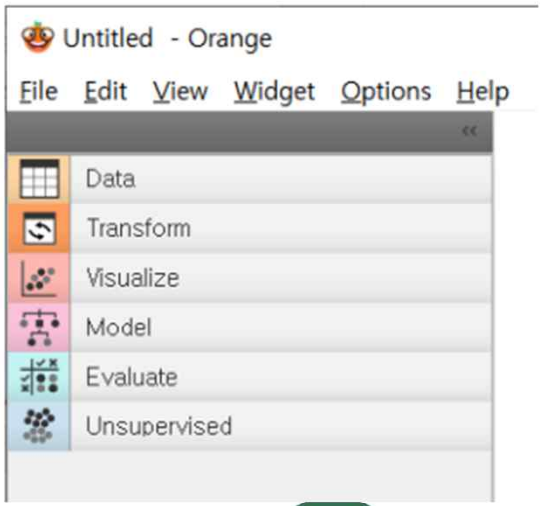


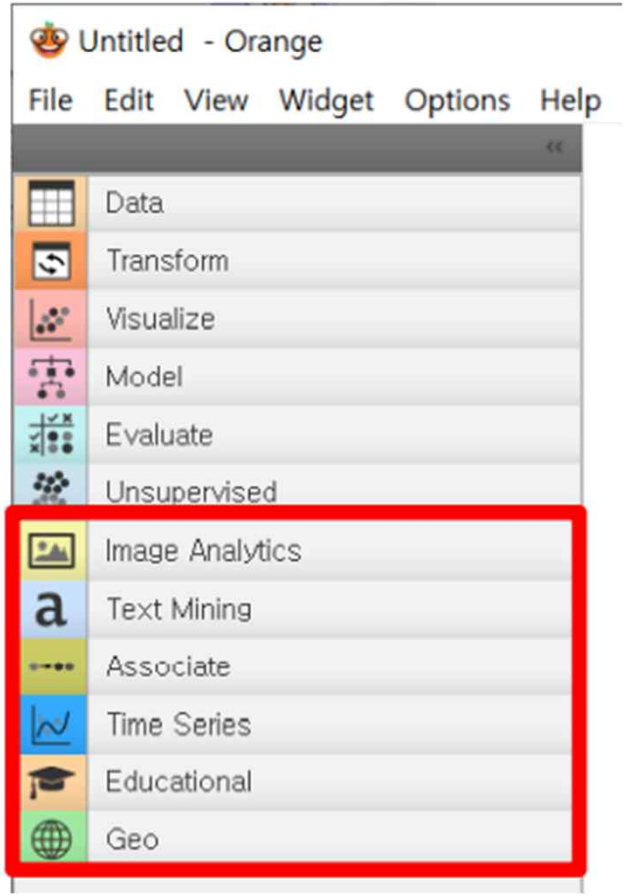
Associate, Educational, Geo, Image Analystics, Text, Timeseries 6개를 체크하고 OK를 누릅니다

Orange3 Add-on 설치

오렌지가 종료되었다가 자동으로 재실행된 후 새로운 위젯이 설치되었음을 확인할 수 있습니다.







다음 시간에는 오렌지에서의 데이터 흐름과 위젯의 활용에 대해 알아보고 정형데이터를 활용한 회귀 분석을 통해 회귀분석의 개념과 절차에 대해 공부해 보도록 하겠습니다. 감사합니다.