청년을 응원합니다!
K-Digital Training

비즈니스 애널리틱스 II

강사 김경하

고용노동부 | 삼정 KPMG | 현대경제연구원 | i-Growth

강의 내용

구분	주제	내용
Part1	빅데이터 개요	데이터 시장 전망, 데이터 엔지니어링의 필요성 데이터엔지니어, 데이터분석가, 데이터사이언티스트
Part2	데이터 분석, 시각화 라이브러리	Ndarray 이해, Numpy 객체 다루기
		Pandas 시리즈, 데이터 프레임, 전처리 방법
		Matplotlib, seaborn, plotly 등 시각화
		실전 분석 프로젝트 도전
Part3	데이터 수집	동적, 정적 웹 데이터 수집, 공공데이터 데이터 활용

데이터 분석 라이브러리

- numpy
- pandas
- matplotlib, seaborn, plotly

numpy 기본기 다지기

Numpy 학습목표

- ndarray 구조와 axis를 안다.
- ndarray의 다양한 슬라이싱 기능을 안다.
- arange 함수를 활용해 ndarray 객체를 만든다.
- ndarray 객체의 데이터를 처리하는 함수를 한다.
- numpy 행렬연산 기능을 알고 활용할 수 있다.
- numpy 기본기 다지기

Numpy 모듈 소개

수학, 과학 계산용 패키지

import numpy

import numpy as np

공식문서 : https://numpy.org/
디지털 도서 : http://bigdata.Dongguk.ac.kr/lectures/Python/_book/numpy.html

Numpy 모듈 소개

- **Num**erical **Py**thon
- 파이썬 내장함수인 리스트보다 **데이터의 저장 및 처리가 효율적**임
- n dimension array(n 차원 배열) - ndarray 타입 제공
- 선형대수와 관련된 기능 제공(**행렬연산**)
- 대규모 다차원 배열을 쉽게 처리하는 기능 지원
- 데이터 과학 도구의 핵심 패키지
- Numpy 기반 패키지 : pandas, scipy, scikit-learn 패키지 등

Numpy 배열 객체(ndarray)

- Numpy 배열 객체 기능
  - 단일요소에 접근하는 인덱싱
  - 하위배열에 접근하는 슬라이싱
  - Bool 배열을 이용한 마스킹 연산
  - 인덱스 배열을 이용한 팬시 인덱싱
  - 모든 기능을 결합한 복합 인덱스 기능 제공
- 브로드캐스팅 기능을 이용해 벡터화 연산 지원

* [참고] 벡터화 연산 지원 : 벡터의 같은 인덱스에 위치한 원소들끼리 연산 수행

행열 데이터 구조

출처 : https://statkclee.github.io/deep-learning/fig/tensor.jpg

Numpy array(배열)

- 여러 값들의 모음, 한 **array 안에 저장 type은 동일**해야 함.

1	2	3	4

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

1차원 배열

numpy.array([1,2,3,4])

2차원 배열

```
numpy.array([[1,2,3,4],
             [5,6,7,8],
             [9,10,11,12],
             [13,14,15,16]])
```

Numpy - matrix(행렬)

Matrix : 행렬

행(row) + 열(column)

ndarray - shape

- ndarray 차원 속성 확인 : arr.shape

  n dimension array(n 차원 배열)

(3,) : 3x1의 배열 : 1차원 배열

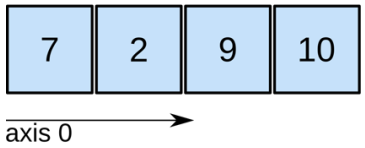
(4,3): 4x3의 배열 : 2차원 배열

(2,5,3): 2x5x3의 배열 : 3차원 배열

axis : 기준이 되는 축

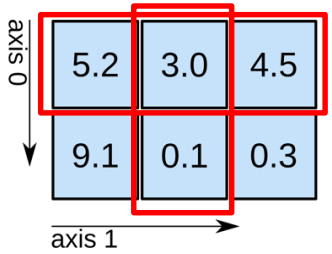
- 1차원 축(행) : axis 0
- 2차원 축(열) : axis 1
- 3차원 축(채널) : axis 2

1D array



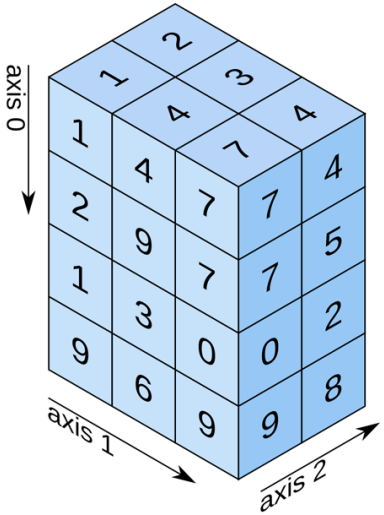
shape: (4,)

2D array



shape: (2, 3)

3D array



shape: (4, 3, 2)

Numpy - matrix(행렬) 연산

- transpose(전치행렬)

  np.transpose(matrix_data)
- 덧셈(+), 뺄셈(-), 위치 곱셈(*)
- dot product

  np.dot(a, b)

  a.dot(b)

Numpy - matrix(행렬) 연산

- 덧셈(+), 뺄셈(-), 위치 곱셈 연산자(*)
- **Shape이 같아야 함**, 같은 위치 끼리 연산

1	2	3
2	3	4

(2, 3)

+

3	4	5
6	7	8

(2, 3)

=

1+3	2+4	3+5
2+6	3+7	4+8

(2, 3)

Numpy - matrix(행렬) 연산

- 행렬 곱(·), dot product : np.dot(a, b), a.dot(b)

불가능

1	2	3
2	3	4

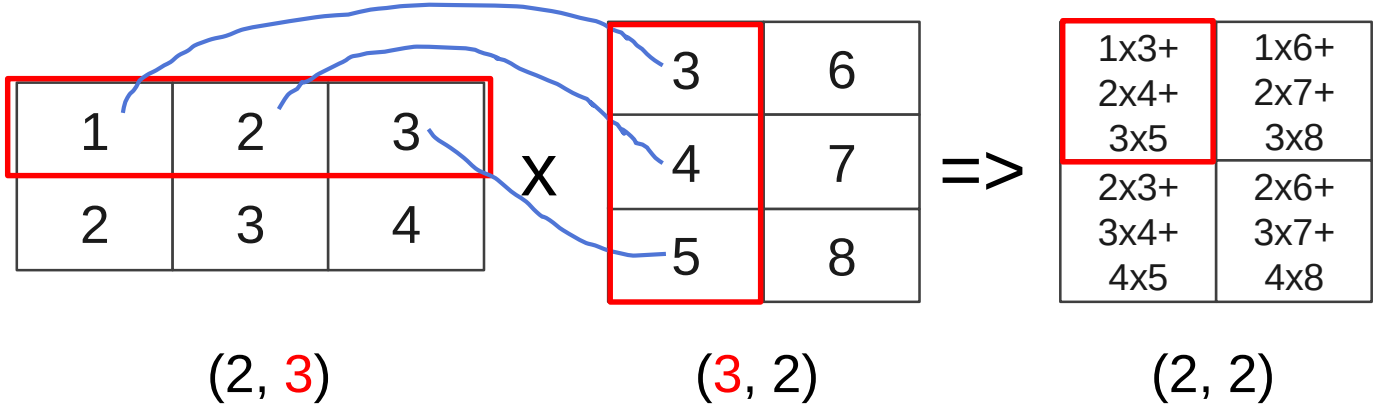
(2, 3)

x

3	4	5
6	7	8

(2, 2)

가능



Numpy - matrix(행렬) 연산

- transpose(전치행렬) : np.transpose(matrix)

1	2	3
5	6	7

(2, 3)

$(matrix)^T$

1	5
2	6
3	7

(3, 2)

Numpy 실습

- 1.a 강의 Numpy 실습

pandas 기본기 다지기

Pandas 학습목표

- csv 파일을 로딩하여 DataFrame을 만든다.
- Series와 DataFrame의 구조를 안다.
- DataFrame의 요약 정보를 확인한다.
- NaN 결측 데이터를 확인하고 다른 값으로 대체한다.
- DataFrame의 인덱싱, 슬라이싱 하여 정보를 추출한다.
- concat, merge 등 다양한 DataFram 조작 함수를 사용한다.

- pandas 기본기 다지기(데이터 전처리를 위한)

공식 문서 : https://pandas.pydata.org/
판다스 기본서 : https://wikidocs.net/book/7188
판다스 300제 : https://wikidocs.net/book/4852

Pandas 소개

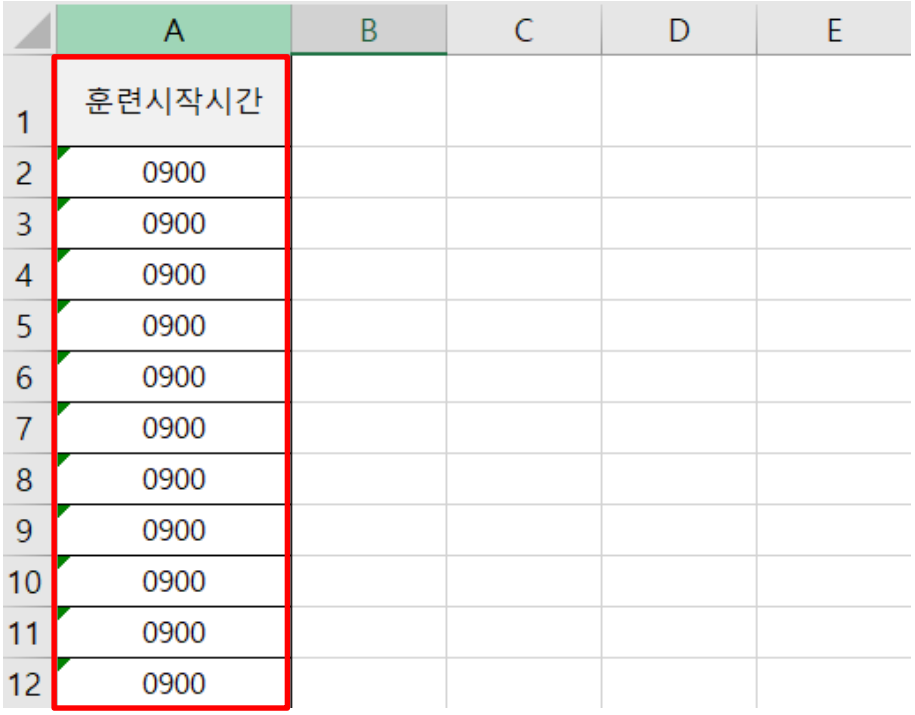
2D 데이터 처리, 막강한 기능 제공

- 데이터를 수집하고 정리하는데 최적화 되어 있음
- 엑셀로 할 수 있는 모든 것들
- csv, excel파일, DB파일, pdf 읽기 등
- 크롤링(웹 정보 수집)
- Database 핸들링
- 시각화

```
import pandas
import pandas as pd
```

Pandas - Series

- Series : 1차원으로 이뤄짐 배열, 1개의 열 Vector



	A	B	C	D	E
1	훈련시작시간				
2	0900				
3	0900				
4	0900				
5	0900				
6	0900				
7	0900				
8	0900				
9	0900				
10	0900				
11	0900				
12	0900				

Pandas - DataFrame

- DataFrame : 2차원으로 이뤄짐 배열

	A	B	C	D	E
1	일자별 시간표 내역을 입력합니다				
2	훈련일자	훈련시작시간	훈련종료시	방학여	시작시간
27	20210902	0900	1800		0900
28	20210902	0900	1800		1200
29	20210902	0900	1800		1300
30	20210903	0900	1800		0900
31	20210903	0900	1800		1200
32	20210903	0900	1800		1300
33	20210906	0900	1800		0900
34	20210906	0900	1800		1200

Pandas 실습

- 1.a 강의 pandas.ipynb
- 1.b[실습1]pandas기본_와인분석.ipynb
- 1.c[실습2]pandas문제해결_와인분석.ipynb

데이터 시각화 하기

- matplotlib
- seaborn

학습목표

- Matplotlib의 차트 구조를 안다.
- Matplotlib 기능을 활용해 데이터를 시각화 한다.
- 차트제목, x축, y축 제목을 표현한다.
- Matplotlib의 다양한 차트를 활용한다.
- seaborn 라이브러리를 활용해 다양한 차트를 표현한다.

- **데이터 시각화의 기본기 다지기**

시각화의 필요성

- 15 x 8 데이터 ; 소량의 데이터

	이름	그룹	소속사	성별	생년월일	키	혈액형	브랜드평판지수
0	지민	방탄소년단	빅히트	남자	1995-10-13	173.6	A	10523260
1	지드래곤	빅뱅	YG	남자	1988-08-18	177.0	A	9916947
2	강다니엘	NaN	커넥트	남자	1996-12-10	180.0	A	8273745
3	뷔	방탄소년단	빅히트	남자	1995-12-30	178.0	AB	8073501
4	화사	마마무	RBW	여자	1995-07-23	162.1	A	7650928
5	정국	방탄소년단	빅히트	남자	1997-09-01	178.0	A	5208335
6	민현	뉴이스트	플레디스	남자	1995-08-09	182.3	O	4989792
7	소연	아이들	큐브	여자	1998-08-26	NaN	B	4668615
8	진	방탄소년단	빅히트	남자	1992-12-04	179.2	O	4570308
9	하성운	핫샷	스타크루이엔티	남자	1994-03-22	167.1	A	4036489
10	태연	소녀시대	SM	여자	1989-03-09	NaN	A	3918661

데이터로 부터 인사이트 얻기가 어렵지는 않다.

시각화의 필요성

- row 307511 x 122 column ; 대량의 데이터

train

	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	FLAG_OWN_CAR	FLAG_OWN_REALTY	CNT_CHILDREN	AMT_INCOME_TOTAL	A
0	100002	1	Cash loans	M	N	Y	0	202500.0	
1	100003	0	Cash loans	F	N	N	0	270000.0	
2	100004	0	Revolving loans	M	Y	Y	0	67500.0	
3	100006	0	Cash loans	F	N	Y	0	135000.0	
4	100007	0	Cash loans	M	N	Y	0	121500.0	
...	...	...	...	...	...	...	...	...	
307506	456251	0	Cash loans	M	N	N	0	157500.0	
307507	456252	0	Cash loans	F	N	Y	0	72000.0	

데이터의 인사이트 도출, 특징을 파악하기 어렵다.

시각화의 필요성

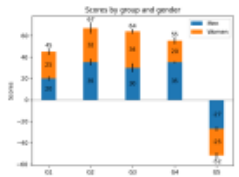
수 많은 숫자 + 문자형으로 된 데이터로 부터

사람이 이해하기 쉽도록 함

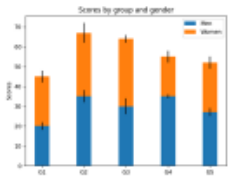
사람이 인사이트를 얻을 수 있는

가장 직관적인 방법

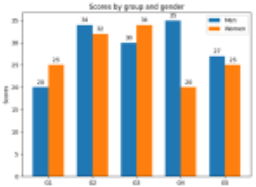
다양한 시각화 방법들



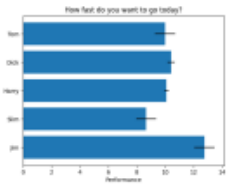
Bar Label Demo



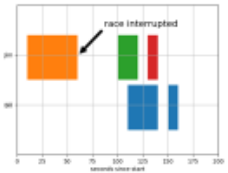
Stacked bar chart



Grouped bar chart with labels

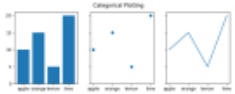


Horizontal bar chart

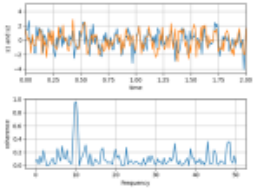


Broken Barh

CapStyle

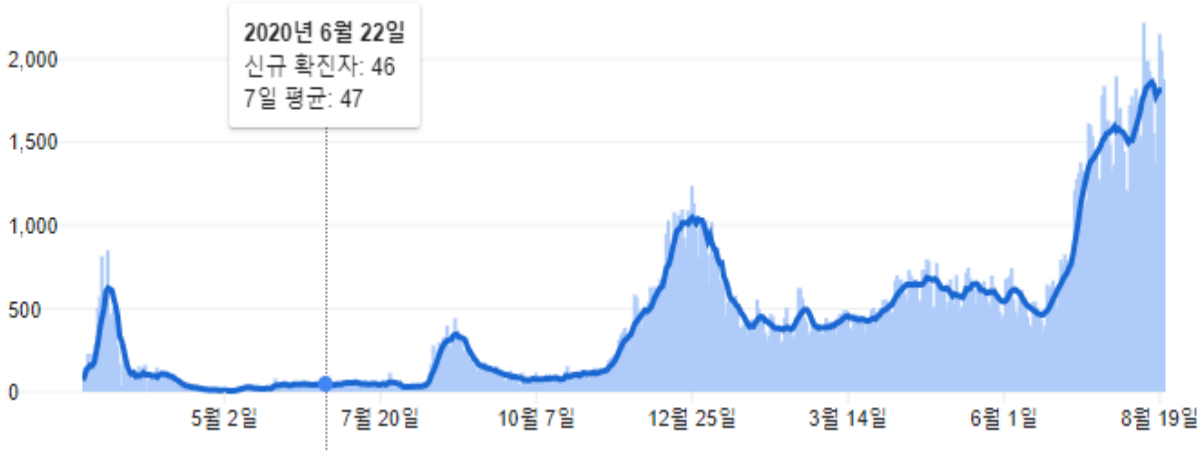


Plotting categorical variables



Plotting the coherence of two signals

효율적인 시각화



	Province/State	Country/Region	Lat	Long	1/22/20	1/23/20	1/24/20	1/25/20	1/26/20	1/27/20	1/28/20	1/29/20	1/30/20	1/31/20	2/1/20	2/2/20	2/
2		Thailand	15.0	101.0	2	3	5	7	8	8	14	14	14	19	19	19	19
3		Japan	36.0	138.0	2	1	2	2	4	4	7	7	11	15	20	20	20
4		Singapore	1.2833	103.8333	0	1	3	3	4	5	7	7	10	13	16	18	18
5		Nepal	28.1667	84.25	0	0	0	1	1	1	1	1	1	1	1	1	1
6		Malaysia	2.5	112.5	0	0	0	3	4	4	4	7	8	8	8	8	8
7	British Columbia	Canada	49.2827	-123.1207	0	0	0	0	0	0	1	1	1	1	1	1	1
8	New South Wales	Australia	-33.8688	151.2093	0	0	0	0	3	4	4	4	4	4	4	4	4
9	Victoria	Australia	-37.8136	144.9631	0	0	0	0	1	1	1	1	2	3	4	4	4
10	Queensland	Australia	-28.0167	153.4	0	0	0	0	0	0	0	1	3	2	3	2	2
11		Cambodia	11.55	104.9167	0	0	0	0	0	1	1	1	1	1	1	1	1
12		Sri Lanka	7.0	81.0	0	0	0	0	0	1	1	1	1	1	1	1	1
13		Germany	51.0	9.0	0	0	0	0	0	1	4	4	4	5	8	10	12
14		Finland	64.0	26.0	0	0	0	0	0	0	0	1	1	1	1	1	1
15		United Arab Emirates	24.0	54.0	0	0	0	0	0	0	0	4	4	4	4	5	5

효율적인 시각화

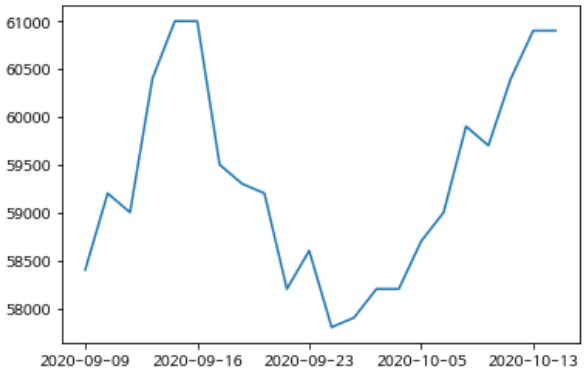

72,700 KRW
−400 (0.55%) 오늘
8월 20일 오후 3:30 GMT+9 · 면책조항
KRX: 005930
+ 팔로우
1일
5일
1개월
6개월
연중
1년
5년
최대
73,600
73,400
73,200
73,000
72,800
72,600
전일 종가 73,100
73,600 KRW 오전 9:45
오전 10:00
오후 1:00
오후 3:00
시가 73,500
최고 73,900
최저 72,500
시가총액 490.27조
주가수익률 15.31
배당수익률 4.14%
전일 종가 73,100
52-주 최고 96,800
52-주 최저 54,000

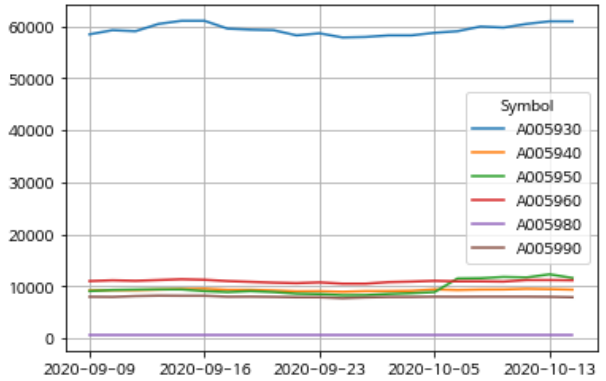
데이터 시각화 종류

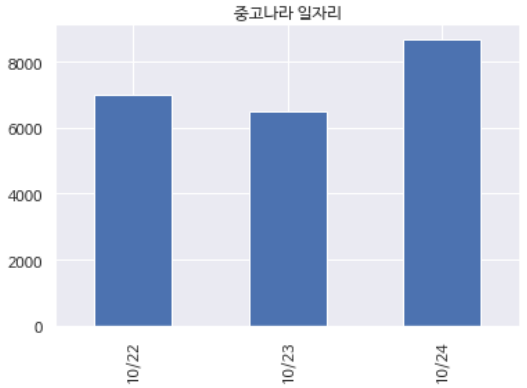
종류	내용
시간시각화	- 분절형과 연속형 - 특정시점 또는 특정시간의 구간 값을 막대 그래프, 누적막대 그래프, 점 그래프 등으로 표현
분포 시각화	- 전체분포와 시간에 따른 분포 - 전체 분포 : 파이차트, 도넛차트, 누적막대, 박스플롯 그래프 - 시간에 따른 분포 : 누적연속 그래프, 누적영역그래프, 선그래프
관계 시각화	- 상관관계, 분포, 비교 - 각 기 다른 변수 사이에서 관계를 시각화 - 스캐터플롯, 행렬, 버블차트 등으로 표현
비교 시각화	- 히트맵, 아웃라이어 찾기(box 플롯)

시간적 시각화(시계열)

- 경제활동 : 국내 총생산, 소비자 물가지수, 수출액
- 물리적활동 : 일일강수량, 기온, 습도
- 인구관련 : 총인구, 농가 수
- 사회생활과 관련 : 교통사고 건수, 범죄발생 수
- 품질 생산관리 등 : 품질 지수, 생산성

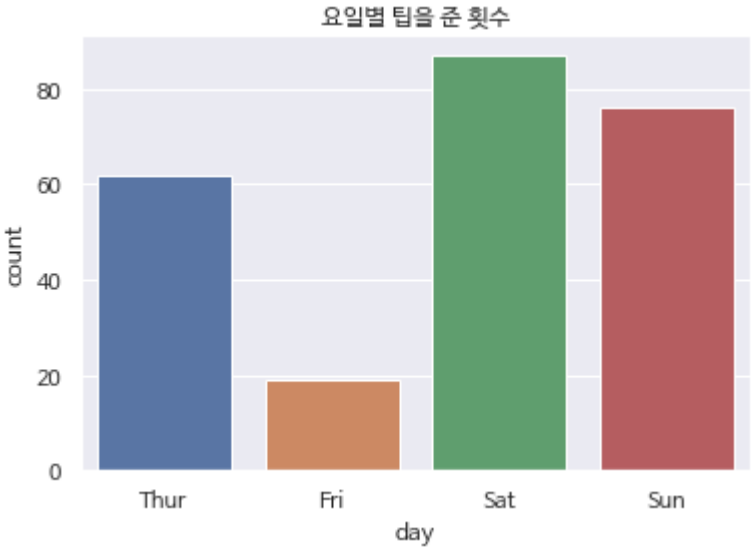


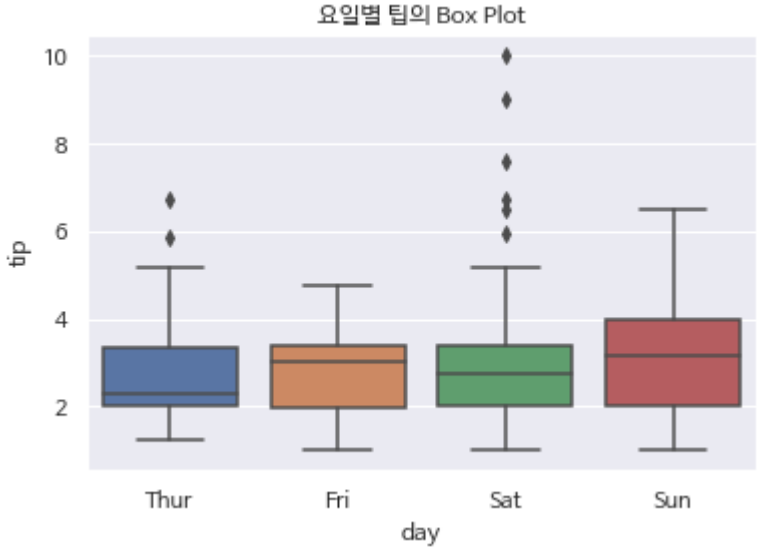




분포 시각화 이해

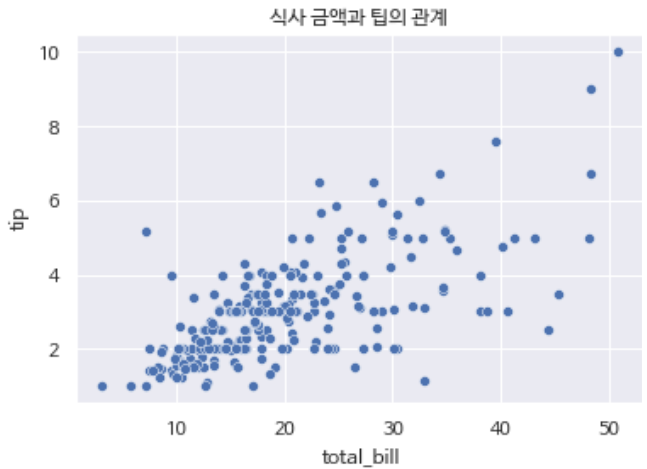
- 분포 데이터 구분 : 샘플 측정 범위에서의 분류
- 분포 데이터 특성 : 최대, 최소, 전체 분포로 나눔
- 전체에 대한 데이터의 양이나 크기를 표현 할 때 사용

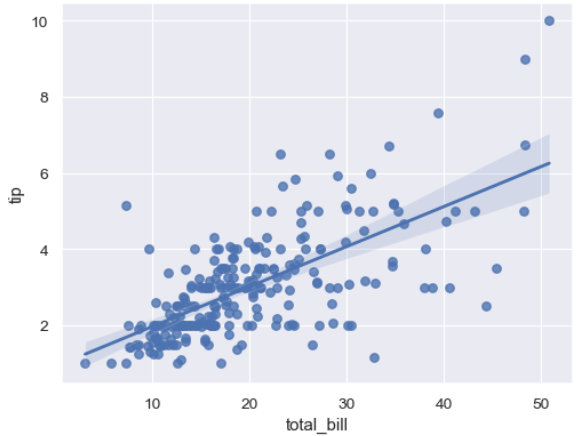


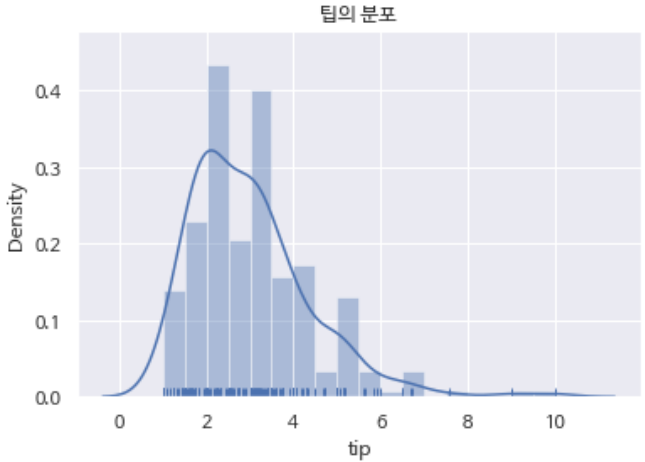


관계 시각화

- 어떤 항목이 다른 항목에 어떤 영향을 주는지 알기 위해 사용
- Scatter 변수 간의 관계를 설명하기 위해
- Histogram : 측정값을 몇 개의 구간으로 나누어 표현한 차트
- Bubble Chart : 스캐터 플롯 + 버블의 크기

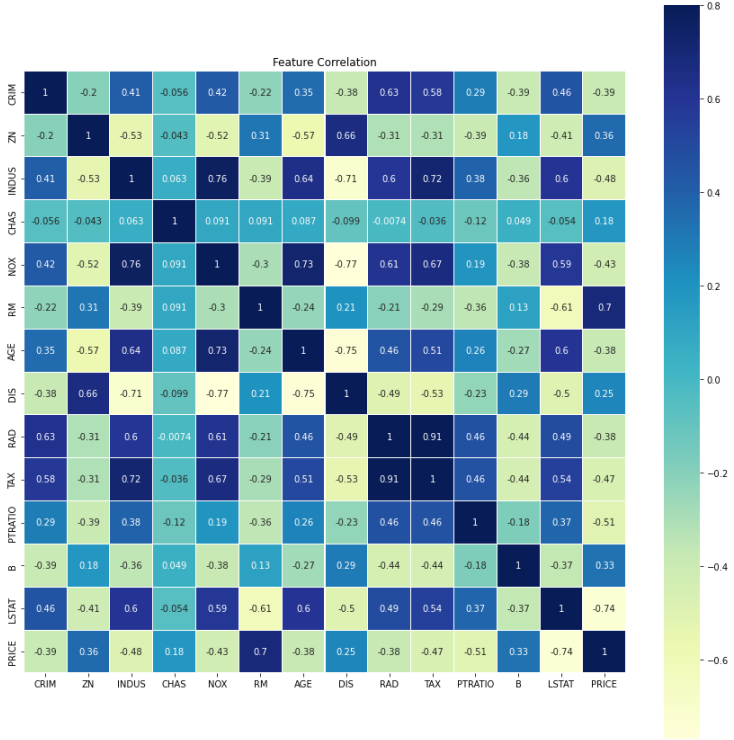






비교 시각화

- 다양한 변수의 특징을 한 번에 비교하여 전체적인 정보 표현이 가능함
- 히트맵 : 색상의 명암으로 값의 크기 표현한 차트



파이썬 시각화 도구

- Pandas의 내장 plot
- Matplotlib
  - 그래프를 그리거나, 분포를 보여주는 파이썬 시각화 패키지
  - 연구용으로 많이 쓰인 MATLAB의 코드 스타일 모방
  - https://matplotlib.org
- Seaborn
  - matplotlib을 wrapping하여 만든 보다 쉬운 패키지
  - 다양하고 화려한 그래프 제공
  - matplotlib 보다 쉽고 단순한 코드
  - https://seaborn.pydata.org
- Plotly
  - 클라우드 서비스, 로컬은 무료로 사용
  - https://plotly.com/python

Matplotlib

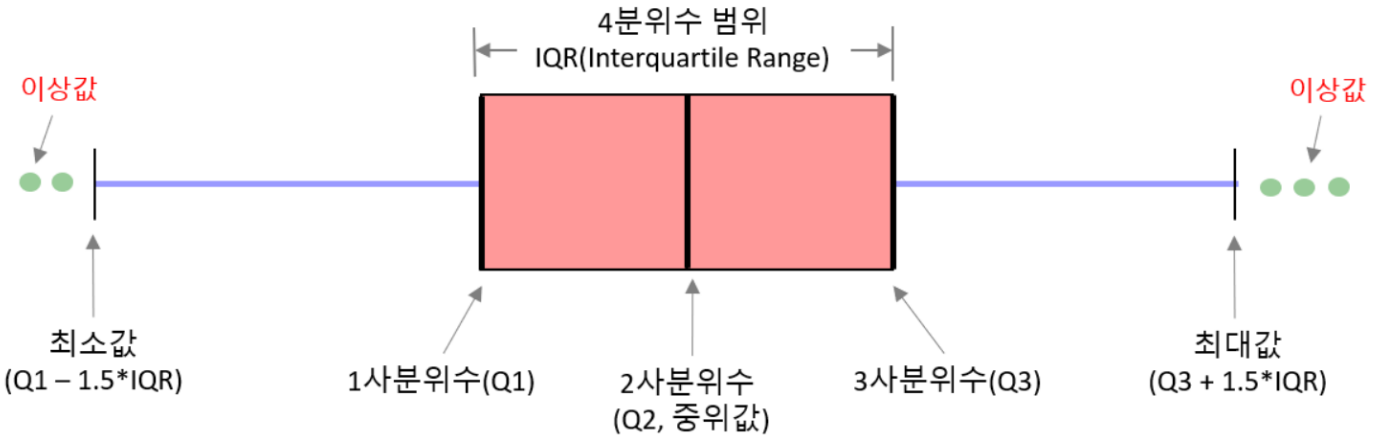
Matplotlib 특징

- 장점
  - 파이썬 표준 시각화 도구로 다양한 기능 지원
  - 세부 옵션을 통해 좀더 예쁜 스타일링 가능
  - 보다 다양한 그래프를 그릴 수 있음.
  - pandas와 연동이 용이함
- 단점
  - 한글 지원이 완벽하지 않음.
    - jupyterlab, colab에서 한글 사용시 추가 설정 필요
  - 세부 기능이 많으나, 사용성이 복잡함

공식문서 : https://matplotlib.org/stable/gallery/index.html
디지털 도서 : https://wikidocs.net/124976

박스플롯 살펴보기

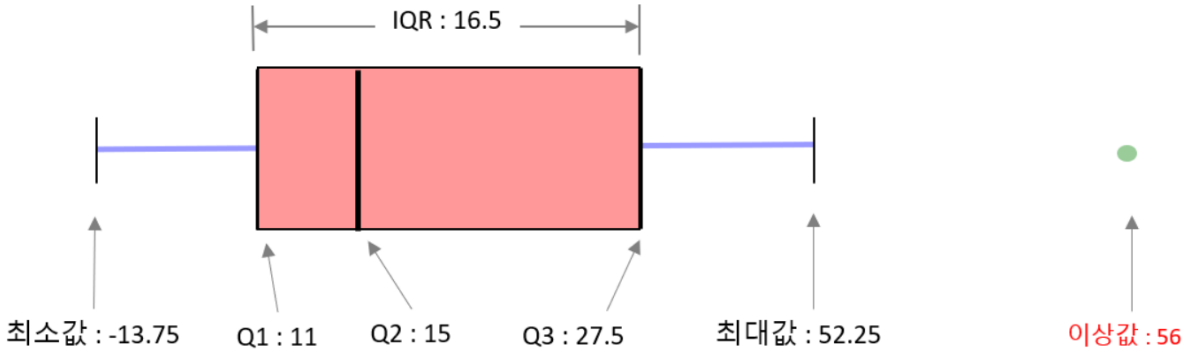
- Boxplot
  - 박스 플롯은 관측값의 대략적 분포와 개별적 이상치를 파악할 수 있는 시각화 차트,
  - 한 공간에서 여러 개의 관측값 그룹 시각화



박스플롯 살펴보기

- 관측값 : 1, 6, 10, 12, 12, 15, 21, 22, 33, 37, 56

2사분위수(중앙값):	15
1사분위수	2Q 기준 왼쪽 중앙값: (10+12)/2 =11
3사분위수	2Q 기준 오른쪽 중앙값 : (22+33)/2=27.5
4분위 범위(IQR)	3Q-1Q =27.5-11=16.5
최소값	1사분위수 – (1.5 * IQR) = 11-(1.5*16.5) = -13.75
최대값	3사분위수 + (1.5 * IQR) = 27.5+(1.5*16.5) = 52.25
이상값	최소값보다 작거나 최대값보다 큰값 -> 56



Seaborn

seaborn 특징

- matplotlib 기반 라이브러리
- 장점
  - 보다 예쁨
  - 통계기능 기반 차트(countplot, replot, lmplot)
  - 쉬운 사용성
  - pandas, matplotlib 호환

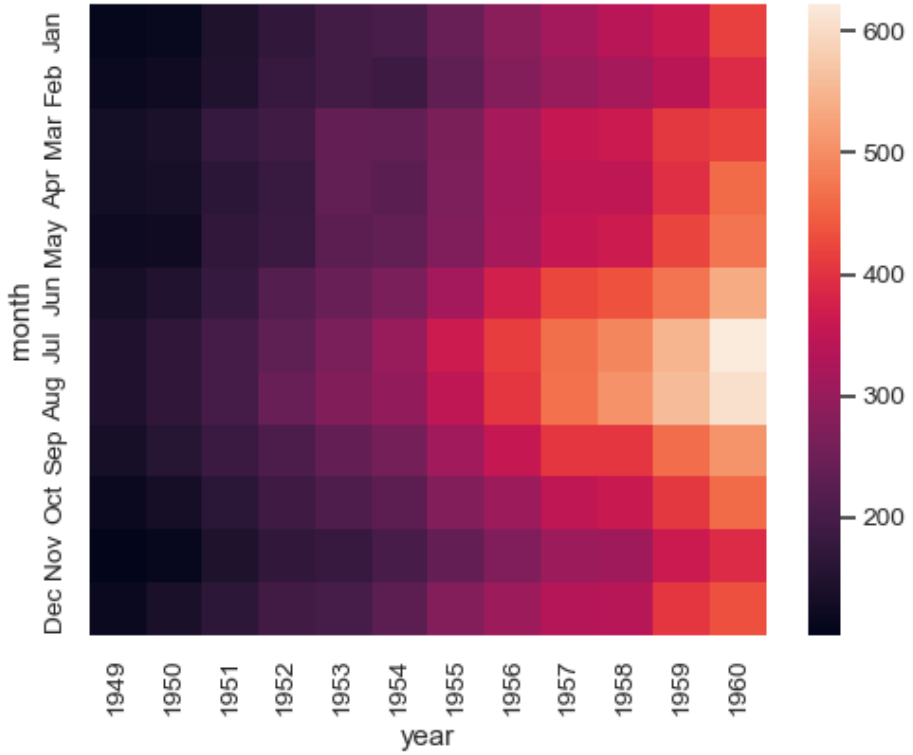
공식문서 : https://seaborn.pydata.org/
디지털 도서 : https://wikidocs.net/86290

Seaborn 데이터 시각화

- matplotlob과 statsmodel 패키지를 이용하여 만듬
- 데이터의 통계적인 부분을 살펴볼 때, matplotlib에 비해 쉽고 간편함
- 내장 데이터 셋 제공
- 다양한 그래프
  - 데이터의 수를 세는 countplot
  - 4분위 도표 boxplot
  - 두 변수 관계를 점 찍어 그리는 scatterplot
  - 두 변수 관계를 점과 분포로 보기 jointplot
  - Hue 인자로 x축 안에서 botplot 그래프 분리
  - 바이올린 모양의 그래프 viloinplot
  - 데이터의 분포를 그리는 distplot
  - Regression 을 표현하는 regplot

sns.Heapmap()

Seaborn 내장 데이터 참고 ttps://seaborn.pydata.org/generated/seaborn.load_dataset.html

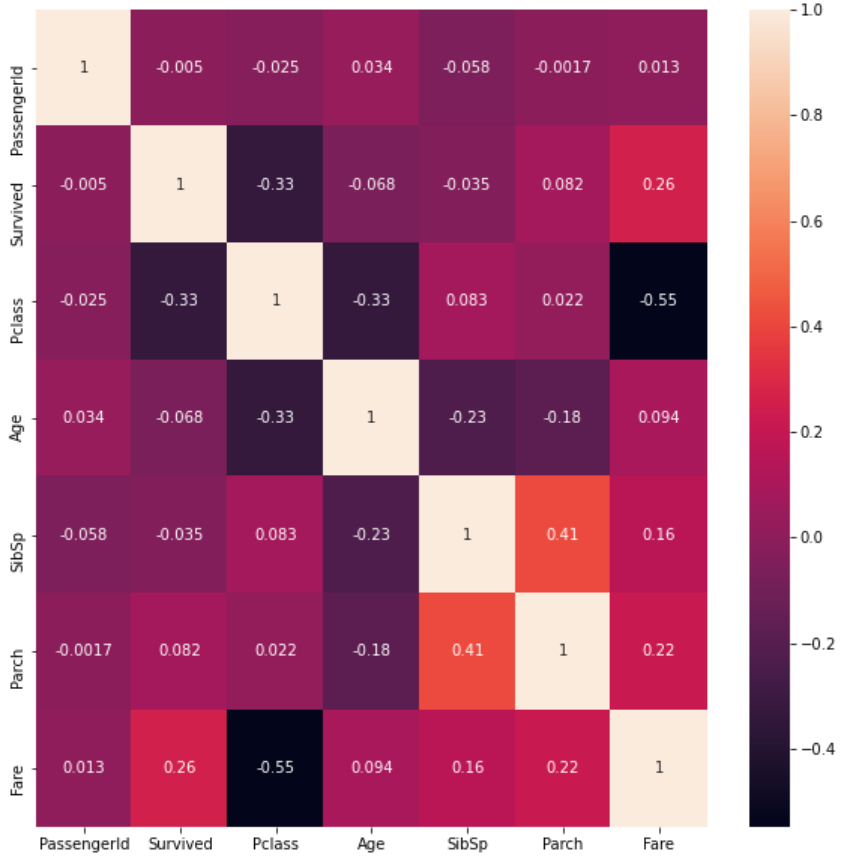


데이터 : 1949~1960년간 승객 수

예) 년도별 월별 승객 수 :

```
flights_df = sns.load_dataset("flights")
flights_df.pivot(
    columns="year",
    index="month",
    values="passengers")
sns.heatmap(flights_df)
```

Heatmp에 상관분석(Correlation Analysis) 적용



sns.heatmap(df_eda.corr(), annot=True)

- 상관분석(상관관계)는 두 변수 간에 어떤 선형적 또는 비선형적 관계를 갖고 있는지 알아보는 방법
- 1에 가까운 값 : 두 변수들 간의 양의 상관관계가 있음.
- 0에 가까운 값 : 두 변수들 간의 상관관계가 없음.
- -1에 가까운 값 : 두 변수들 간의 음의 상관관계가 있음.

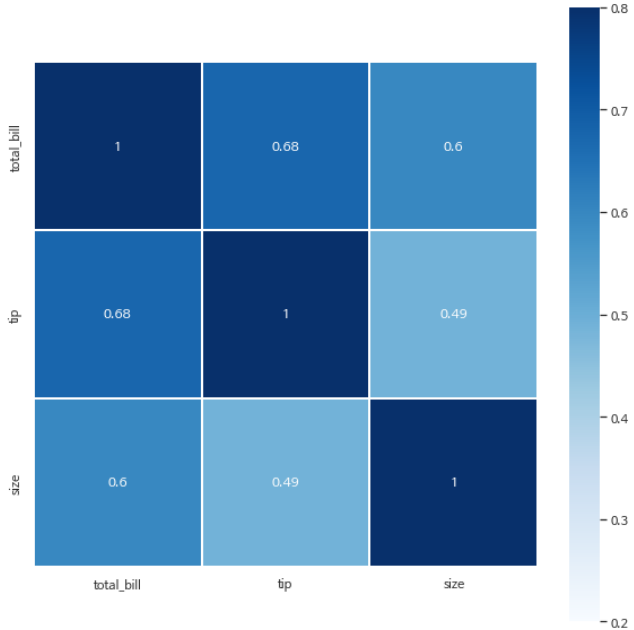
Seaborn heatmap 그리기

- 컬럼 간의 상관관계를 만들어 주는 함수

  corr = pandas.corr()

- searborn의 heatmap()으로 시각화

```
sns.heatmap(corr,   #데이터
    vmin = 0.2,       #최소값 지정(default -1)
    vmax = 0.8,       #최댓값 지정(default 1)
    cbar = True,      #corr color bar 표시
    linewidths=0.01, #cell사이에 선 넣기
    annot = True,     #cell 에 값 표시
    cmap = 'Blues'     #color map 지정
)
```



공식문서 참조:
https://seaborn.pydata.org/generated/seaborn.heatmap.html?highlight=heatmap#seaborn.heatmap

Plotly

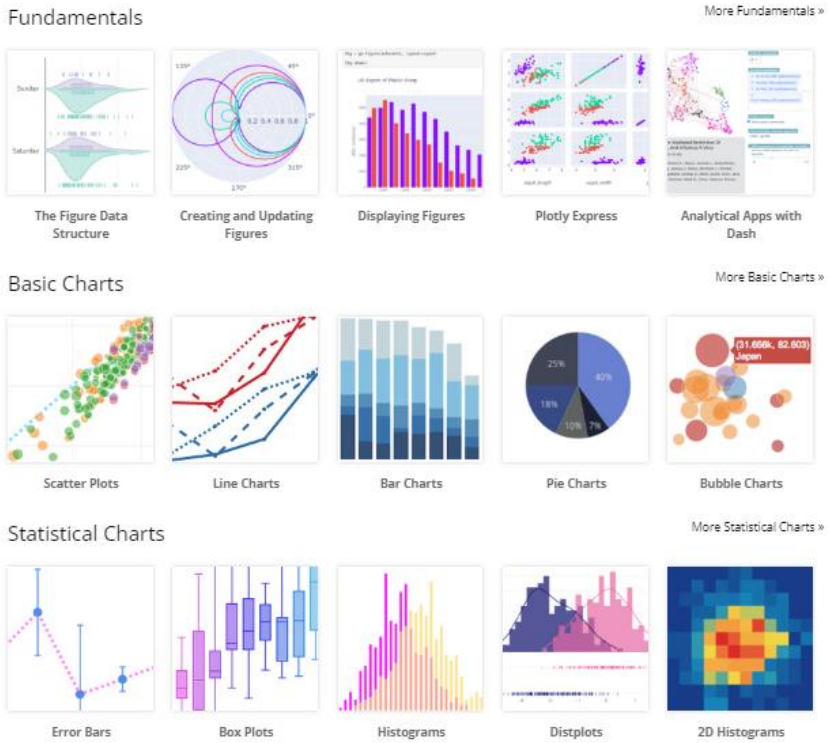
Plotly 소개

- 인터렉티브한 시각화가 가능한 그래픽 라이브러리
- 줌인, 줌아웃 및 툴팁을 활용한 데이터확인이 가능
- Dash, 및 chart Studio 와 같은 visualisation tools 연동으로 WebApp 구현 가능
- matplotlib 대비 코드가 훨씬 간편
- Python, R, Julia, MATLAB 등 여러 프로그래밍 언어 API 제공
- Plotly는 기본적으로 JSON 형태를 주고받는 구조로 되어있음.
- pandas 호환 기능이 추가됨
- 서비스 정책에 따라 무료/유로 존재

공식 문서 https://plotly.com/python/getting-started/
디지털 도서 https://wikidocs.net/book/8909

Plotly 소개

- 기본적인 색감이 매우 이쁨



Plotly 사용

- plotly
  - 설치 : pip install plotly 6.0.1 (2025.04.16 기준)
- ipywidgets 설치(공식문서 참고)
  - https://plotly.com/python/getting-started/#jupyterlab-support
  - 설치 : pip install ipywidgets
- jupyterlab 업그레이드(그래프가 보이지 않으면)
  - 가상환경에 직접 설치를 해야함.
  - pip install --upgrade jupyterlab

Plotly 문서

- 공식 문서 : https://plotly.com/python/
- 디지털 문서 : https://wikidocs.net/185049

Plotly 차트 그리기

- 차트 그리는 2가지 방법

	장점	단점
graph_objects	그래프를 세세하게 구성 가능	코드가 길고 문법이 복잡하여 시간이 오래 걸리고 학습하기 어렵다
express	간단한 코드로 쉽게 그래프 표현가능	정해진 템플릿 외세세한 조정이 어려워 세세한 조정 필요시 graph_objects 도움이 필요함

- press : 사용자가 빠르게 데이터 분석을 진행할 때 활용 추천함
- graph_objects 는 논문, 발표자료와 같이 그래프 visualization에 중점을 두었을 경우 활용 추천함
- 두가지 융합해서도 가능함

Plotly 그래프 그리기

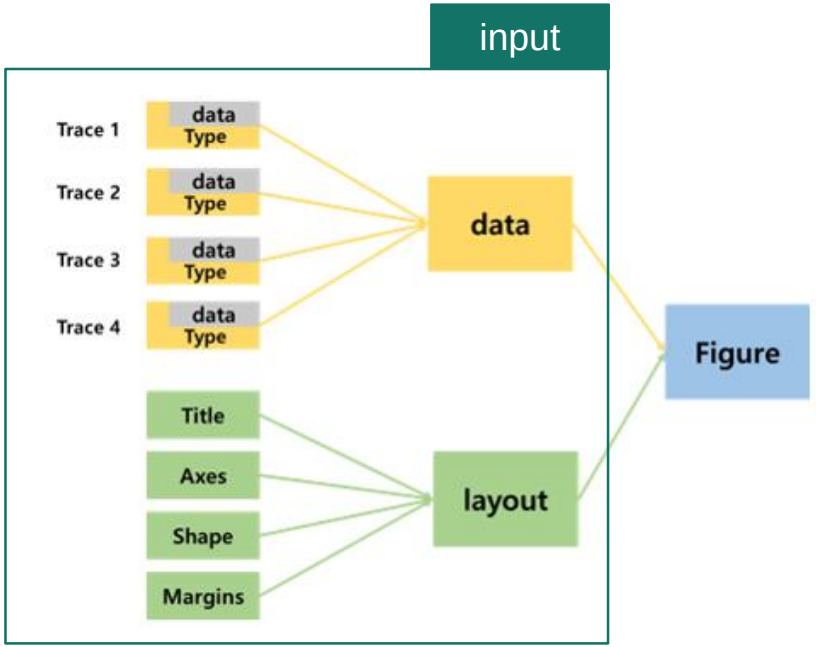
- 1단계 : 기초 그래프 생성
- 2단계 : 다양한 방법으로 그래프를 업데이트함

- Express
- graph_objects

- add_trace()
- update_traces()
- update_layout()
- update_xaxes()
- update_yaxes()

Plotly 그래프 그리기

- Figure는 Ploty 작업의 기본 단위
- go.Figure() 함수를 통해 생성 가능
- go.Figure() 함수의 구조
  - Data
  - Layout



WordCloud로 시각화 하기

WordCloud 시각화

- 참고 교재
  - https://wikidocs.net/172882
- 설치 라이브러리
  - pip install wordcloud
  - pip install konlpy
- java 다운로드 설치하기
  - Konlpy 실행을 위해 필요함
  - https://www.java.com/en/download/
- 대상 텍스트(말뭉치) 위치
  - https://konlpy.org/ko/latest/index.html
  - https://konlpy.org/ko/latest/data/

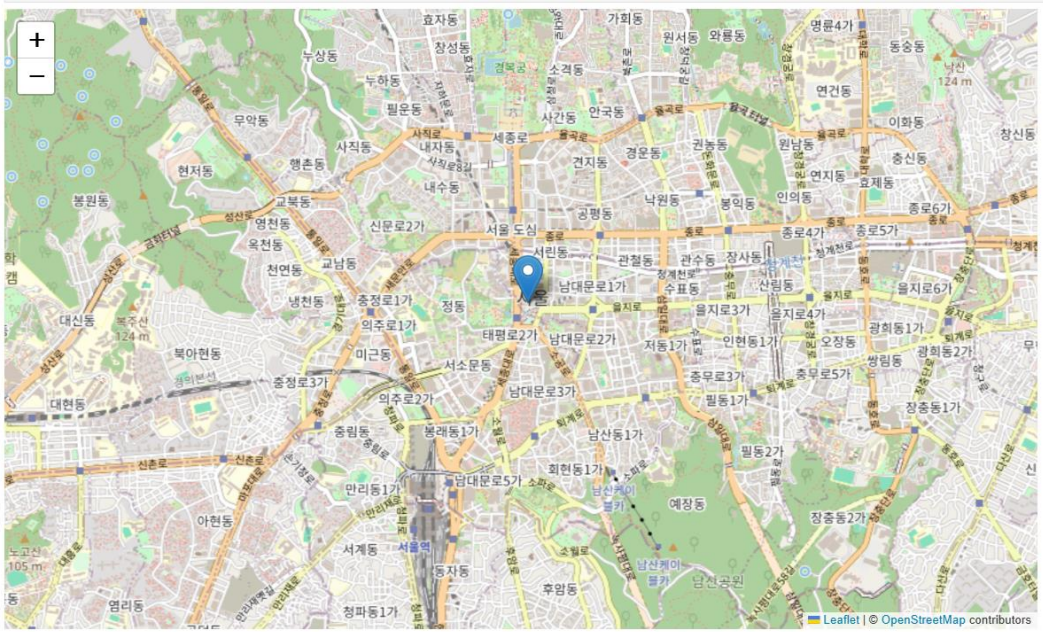
WordCloud 시각화

- font 파일 다운로드
  - curl -L -O https://github.com/byungjooyoo/Dataset/raw/main/NanumGothic.ttf
- 이미지 다운로드
  - curl -L -O https://github.com/byungjooyoo/Dataset/raw/main/korea_map.png
  - curl -L -O https://raw.githubusercontent.com/byungjooyoo/Dataset/main/heart.png

지도로 시각화 하기

폴리엄(Folium) 라이브러리 사용

- 참고 교재
  - https://wikidocs.net/234454
- 라이브러리 설치
  - pip install folium
- 실습 자료
  - 실습 파일



팀별 문제해결

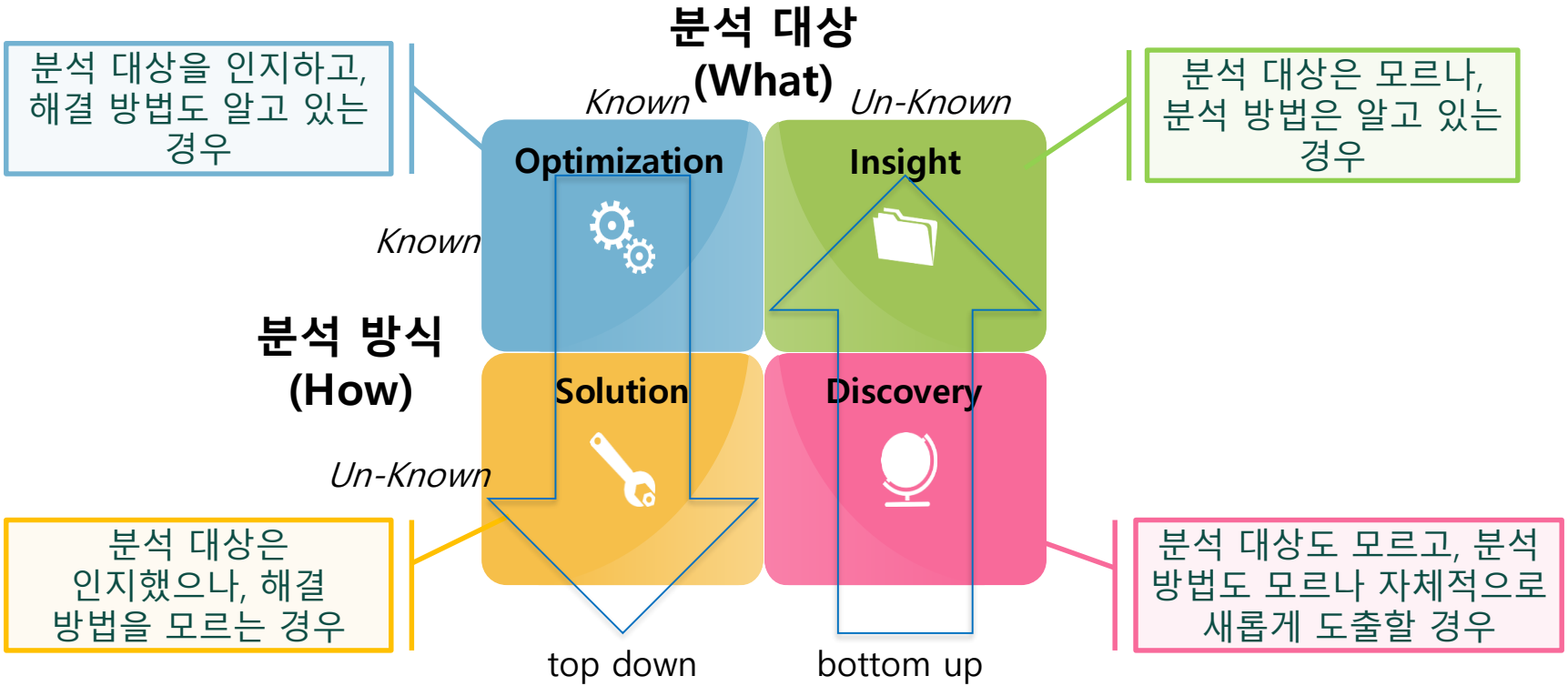
문제해결(팀 미션)

- 우리지역 인구와 연령대별 인구 분포 알아보기
  - 예시 큰 범주
    - 서울시는 구 단위
    - 성남시
    - 수원시
    - 용인시
- 인구 분포를 통해 알 수 있는 행정적 인사이트 찾고 정의하기
- 구청 직원분께 행정정책을 제안한다면 어떻게 할 것 인지 생각해 보기
- 각 동명과 인구를 지도에 표시하기
  - folium또는 다른 지도를 사용해도 됨
  - 사용자가 편하게 마우스를 올리면 툴팁이 표시되도록 하기
- 본인 git 올려서 정리하기

데이터 분석과 머신러닝 모형

데이터 분석에서 가장 중요한 것은?

분석의 대상(What) 및 분석의 방법(How)에 따른 분석 주제 유형



Insight vs. Optimization

Insight

탐색적 자료 분석(Exploratory Data Analysis)

데이터의 특징과 내재하는 구조적인 관계를 알아내기 위한 분석 기법으로 이러한 자료의 탐색 과정을 통하여 얻은 정보를 기초로 통계모형을 세울 수 있음

미지의 특성을 파악하고 자료구조를 파악할 수 있는 증거 수집의 과정

Looking at data to see what it seems to say. It's concentrates on simple arithmetic and easy-to-draw picture. *John Tukey, 1977*

데이터 탐색 및 시각화

의미 있는 정보는 무엇이 있을까?

- 데이터를 잘 다룰 수 있어야 함
- 데이터의 부분집합 추출 및 병합 등의 작업이 주를 이룸
- 데이터 시각화를 통해 탐색 결과를 이해하기 쉽도록 함

EDA를 위한 파이썬 패키지

- Numpy & Pandas : 데이터를 다루기 위한 패키지
- Matplotlib & Seaborn : 데이터를 시각화 하기 위한 패키지

Optimization

기계학습(Machine Learning)

어떻게 하면 더 빨리 학습시키고 어떻게 하면 더 정확히 예측할 수 있을까에 대한 연구를 하며 데이터 전처리, 파생 변수 추가, 모형 선택 등의 방법을 통해서 예측 모형의 평가 점수를 높이는 과정

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E. *Tom Mitchell, 1997*

예측력이 높은 모형 생성

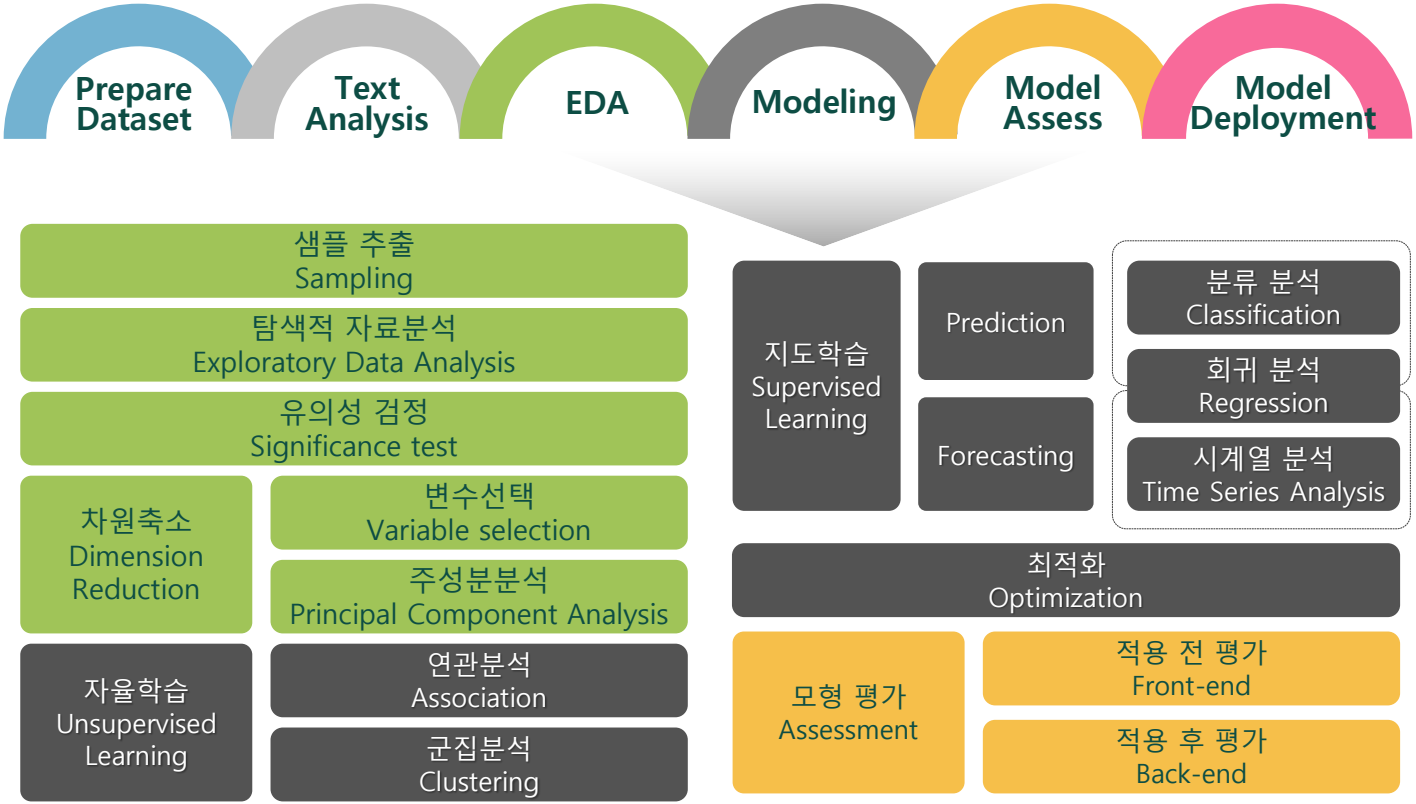
어떻게 하면 모형이 예측을 잘 할 수 있을까?

- 데이터 전처리, 파생변수 추가
- 머신러닝 모형 생성 및 예측
- 모형 평가

ML을 위한 파이썬 패키지

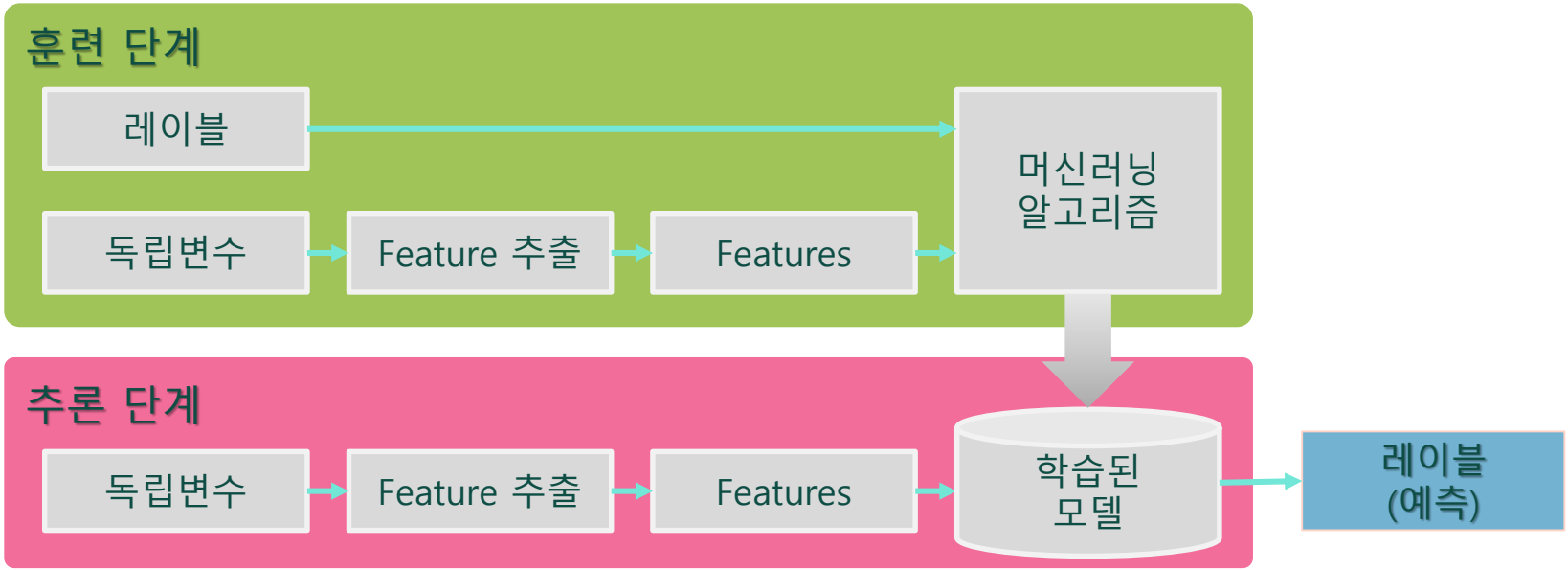
- Scikit-learn : 기계학습 라이브러리
- Statsmodels : 통계 라이브러리

데이터 분석 단계에서 머신러닝



머신러닝 단계

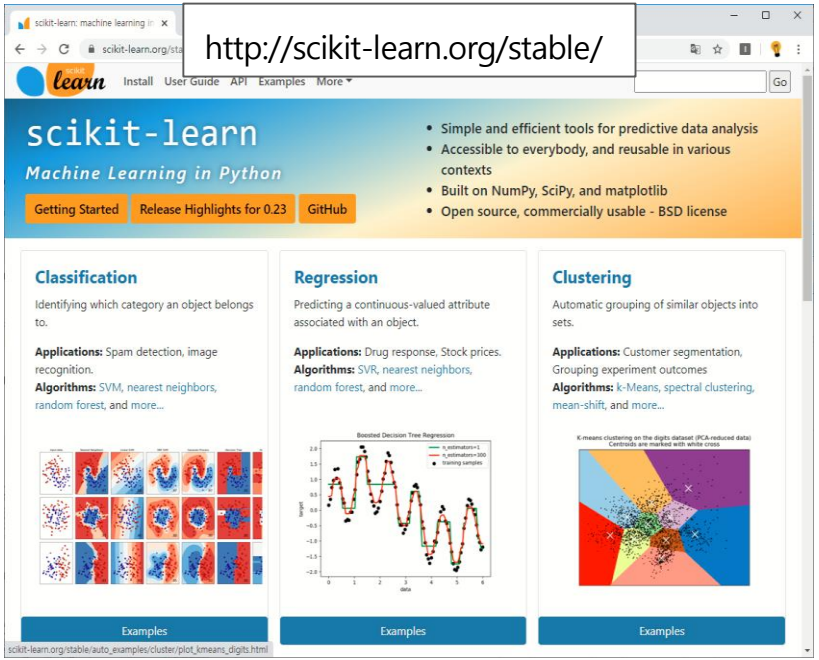
- 학습(training) vs. 추론(Prediction/Inference)
- 학습: 훈련 데이터를 이용하여 모델을 학습하는 과정
- 추론: 학습된 모델을 이용하여 미래의 새로운 데이터를 추론/예측하는 과정



Scikit-learn 패키지

- **예측분석을 위한 간단하고 효율적인 도구**
- **상업적으로 사용 가능한 오픈소스 BSD 라이센스이므로 모든 사람이 사용할 수 있음**
- **NumPy(넘파이), SciPy(사이파이) 및 matplotlib(맷플롯립) 기반**

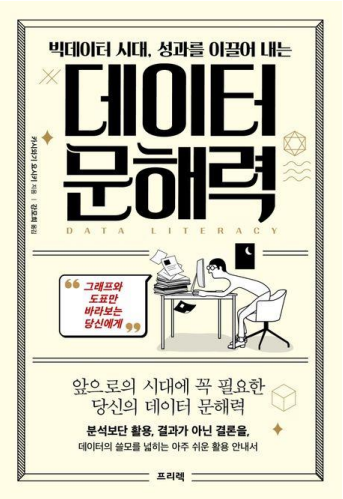
- 사이킷런은 분류(Classification), 회귀(Regression), 군집(Clustering) 분석을 위한 다양한 클래스들이 구현되어 있으며, 이를 통해 예측 모형을 만들 수 있음
- 사이킷런은 차원축소(Dimensionality reduction), 모델 선택(Model selection), 전처리(Preprocessing)를 위한 많은 기능들이 구현되어 있으므로 데이터 분석을 위한 필수 패키지임

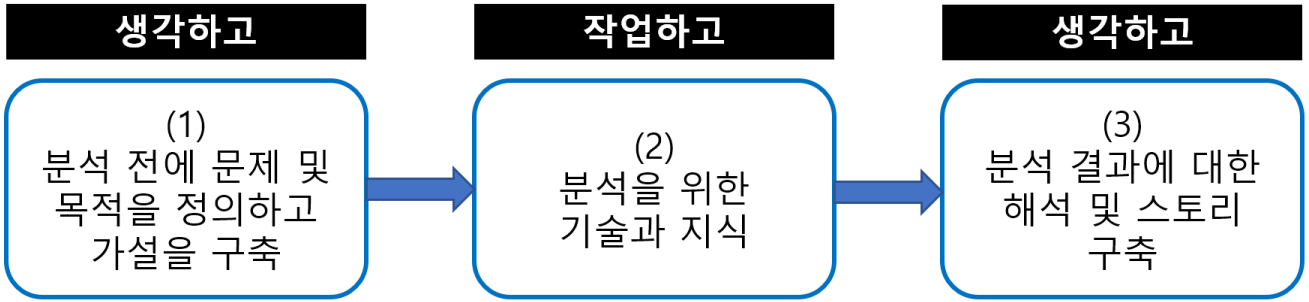


머신러닝 모델은 변환 모델과 예측 모델이 있음

- 학습에 사용하는 함수: fit()
- 변환에 사용하는 함수: transform()
- 예측에 사용하는 함수: predict()

데이터 분석가의 마인드 셋

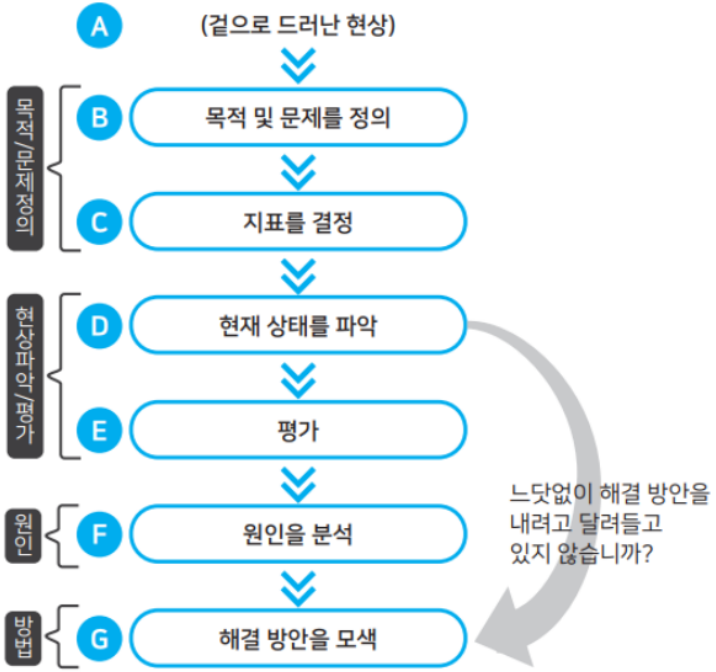




[참고] https://brunch.co.kr/@ashashash/133

데이터 분석가의 마인드 셋

데이터 활용 프로세스 5단계



[참고] https://buly.kr/5fAxNg6

B. 무엇을 알고 싶은지, 어떤 문제를 해결하고자 하는지 구체적이고 명확한 언어로 정리하기

C. 목적과 문제에 대해 논리적인 결론을 낼 수 있는 데이터와 지표를 설정

D. 문제 상황에 대한 데이터를 그래프나 표 등으로 시각화

E. '비교'를 통해 현재 상황에 대해 평가하고 단순히 지표 해석 결과가 아닌 문제 상황에 대한 결론을 도출

F. 문제/결과에 대한 원인을 분석하고 해결방안(목적)을 도출하기 위한 근거를 찾아냄

G. 논리적으로 분석된 원인에 대한 논리적인 해결책을 제시하고 실행

감사합니다