라이브 코칭 3회차

Data Science 2024

잠시 후 오후 8시, 코칭스터디(Data Science 2024) LIVE가 진행됩니다

공지사항

boostcourse

3주차 학습계획/미션

부스트코스 - [스터디강좌]에서 확인하기

필독 | 오리엔테이션

FAQ | 자주묻는 질문

학습 :: 미션 :: 라이브 안내

3주차 | 건강검진 데이터로 가설검정하기

- 3주차 | 학습 계획 & 학습 범위 👈
- 3주차 | 미션 👈
- 3주차 | 라이브 코칭 :: 실시간 시청
- 3주차 | 라이브 코칭 :: VOD 다시보기

💙 💙 3주차 학습 계획 💙

1) 파이썬으로 시작하는 데이터 사이언스 강좌 수강하기(🔽아래의 학습

- (1) 건강검진 데이터로 가설검증하기 강의 듣기
- (2) QUIZ 3 풀기

  ▶▶QUIZ 3 : https://www.boostcourse.org/ds112/quiz/60789

2) 퀴즈 인증 제출하기(🕐08월 18일 일요일 23:59까지!)

- 위의 QUIZ 3 풀이 후 화면 캡쳐해서 슬랙에 업로드하기
- 슬랙(코칭스터디 <Data Science>) → 본인 팀 채널에 업로드 (OO코치_01~10팀)

3) 라이브 코치님께 질문 남기기(💌08월 15일 목요일 18:00까지!)

- 라이브 코치님께 궁금하신 사항이 있으신 분들은 자유롭게 남겨주세요!
- 슬랙(코칭스터디 <Data Science>) → 03-코치에게-물어봐 채널에 남겨주세요!

⏰ 🔔 3주차 미션 내용을 알려드립니다 🔔 ⏰

부스터 여러분들, 3주차 강의는 잘 들으셨나요?!

학습한 내용을 토대로 풀이하여야 할 3주차 미션 내용을 아래와 같이 공개합니다!

미션 내용을 팀원들과 함께 풀이해주세요!(적극적인 토론이 필요합니다!)

📌매주 일요일 23:59까지 리드부스터가 제출해주세요! 모두들 화이팅입니다!

📌미션에 도전하기 전에 먼저!!

!번 미션에 활용되는 데이터를 다운로드 받기 위해, 주피터 노트북에서 다음 셀을 먼저 실행해주세요.

나의 컴퓨터 환경에 데이터를 저장하지 않아도, 웹 사이트에서 바로 데이터를 받아올 수 있습니다.

- 원활한 피드백을 위해 미션을 제출할 때에도 아래 코드를 꼭 포함해서 제출해주세요!
- [참고] 한글폰트 설정 : https://github.com/ychoi-kr/koreanize-matplotlib

```
import pandas as pd
import numpy as np
import seaborn as sns

# 한글폰트 사용을 위해 설치
# 아래 모듈을 설치하고 불러오면 별도의 한글폰트 설정이 필요 없습니다.
# colab 에서도 아래 설치 문구의 주석을 제거하고 설치하고 import 해주기만 하면 한글폰트가 잘 나옵니다.
# !pip install koreanize-matplotlib

import koreanize_matplotlib
```

NAVER CONNECT

모두의연구소

리드부스터 활동일지

8월 18일 일요일 23시 59분까지

boostcourse

퀴즈	1주차	2주차	3주차	4주차	수료 기준: 75% 이상
홍길동_리더	☐	☐	☐	☐	0%
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
미션	1주차	2주차	3주차	4주차	수료 기준: 75% 이상
홍길동_리더	☐	☐	☐	☐	0%
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
라이브	1주차	2주차	3주차	4주차	수료 기준: 75% 이상
홍길동_리더	☐	☐	☐	☐	0%
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	
	☐	☐	☐	☐	

💡 슬랙에서 **과업을 인증**하면 해당 주차에 **체크** ✔ 해주세요.

🚨 **작성하는 공간이 아닙니다**

🚨 **과업 진행률**(자동 표시)

NAVER CONNECT

모두의연구소

2주차 미션 정리

미션 간단 풀이

2주차 미션 솔루션 공유

2주차 미션 출제 의도

1번	[판다스 인덱싱과 연산] 전국 시도별 약국수 구하기
2번	[판다스 인덱싱과 연산] 시도별 동물병원 빈도수 구하기
3번	[판다스 인덱싱과 연산] 피부과나 성형외과가 다른지역에 비해 많은 곳은?
4번	[folium 활용] 지도에서 노인/치매병원 위치 보기

2주차 미션 - 1번

[판다스 인덱싱과 연산] 전국 시도별 약국수 구하기

📌Q1. 상권업종중분류명이 병원인 것을 찾아 빈도수를 구해주세요. 이 때, value_counts, groupby, pivot_table 등 다양한 집계 방법을 통해 구해볼 수 있습니다. 각자 구하기 편한 방법을 통해 빈도수를 구합니다.

```
df.loc[df["상권업종소분류명"] == "약국", "시도명"].value_counts()
```

```
경기도          4510
서울특별시       3579
부산광역시       1130
경상남도         1017
인천광역시       1002
경상북도          915
대구광역시         870
전라북도          862
충청남도          830
전라남도          811
강원도           729
광주광역시         691
충청북도          648
대전광역시         603
울산광역시         362
제주특별자치도      226
세종특별자치시       99
Name: 시도명, dtype: int64
```

2주차 미션 - 2번

[판다스 인덱싱과 연산] 시도별 동물병원 빈도수 구하기

📌 Q2. 여러분은 반려동물과 관련된 사업을 하는 스타트업에 취업을 하여 상권분석을 해달라는 요청을 받았습니다.

병원이나 약국은 인구나 유동인구가 많은 지역에 주로 위치하고 있습니다. 그렇다면 동물병원도 병원이나 약국이 많은 곳에 더 많이 있을까요?

빈도수를 구하고 시각화 하여 동물병원이 어느지역에 많은지 분석해 주세요!

2주차 미션 - 2번

[판다스 인덱싱과 연산] 시도별 동물병원 빈도수 구하기

```
df["상권업종소분류명"].unique()
```

```
array(['산부인과', '내과/외과', '신경외과', '기타병원', '약국', '동물병원', '한약방', '탕제원',
       '정형/성형외과', '소아과', '이비인후과의원', '노인/치매병원', '언어치료', '수의업-종합', '한의원',
       '치과의원', '침구원', '일반병원', '안과의원', '조산원', '한방병원', '종합병원', '유사의료업기타',
       '응급구조대', '혈액원', '치과병원', '척추교정치료', '피부과', '비뇨기과', '치과기공소', '산후조리원',
       '접골원', '수의업-기타', '제대혈'], dtype=object)
```

```
df.loc[df["상권업종소분류명"] == "동물병원", "시도명"].value_counts()
```

```
경기도          992
서울특별시       557
인천광역시       193
경상북도         165
경상남도         161
부산광역시       153
충청남도         131
대구광역시       119
전라북도         111
강원도           85
대전광역시        77
전라남도         77
충청북도         75
광주광역시        71
울산광역시        61
제주특별자치도     46
세종특별자치시     13
Name: 시도명, dtype: int64
```

2주차 미션 - 3번

피부과나 성형외과가 다른지역에 비해 많은 곳은?

📌 Q3. 강남지역에는 다른 지역에 비해 피부과나 성형외과가 많아보입니다.
실제로 해당 지역에 피부과나 성형외과가 다른지역에 비해 전체 병원 수 중에서 어느정도의 비율을 차지하고 있는지 구별로 구해주세요.

상권업종소분류명에 "피부" 나 "성형"이 들어가는 서울시에 소재한 병원을 찾아주세요.
그리고 시군구별로 피부, 성형이 들어가는 비율이 어느정도 되는지 구해주세요.

2주차 미션 - 3번

피부과나 성형외과가 다른지역에 비해 많은 곳은?

```
df["피부성형"] = df["상권업종소분류명"].str.contains("피부|성형")
df_seoul_hospital = df[(df["시도명"] == "서울특별시") & (df["상권업종중분류명"] == "병원")]
df_seoul_hospital.groupby(["시군구명"])["피부성형"].mean().round(2).sort_values(ascending=False)
```

```
시군구명
강남구      0.24
서초구      0.17
마포구      0.09
영등포구     0.08
서대문구     0.08
강서구      0.07
강동구      0.07
중구       0.07
양천구      0.07
중랑구      0.07
성북구      0.06
송파구      0.06
구로구      0.06
은평구      0.06
광진구      0.06
노원구      0.05
성동구      0.05
금천구      0.05
용산구      0.05
강북구      0.05
동대문구     0.04
동작구      0.04
관악구      0.04
종로구      0.03
도봉구      0.02
Name: 피부성형, dtype: float64
```

2주차 미션 - 4번

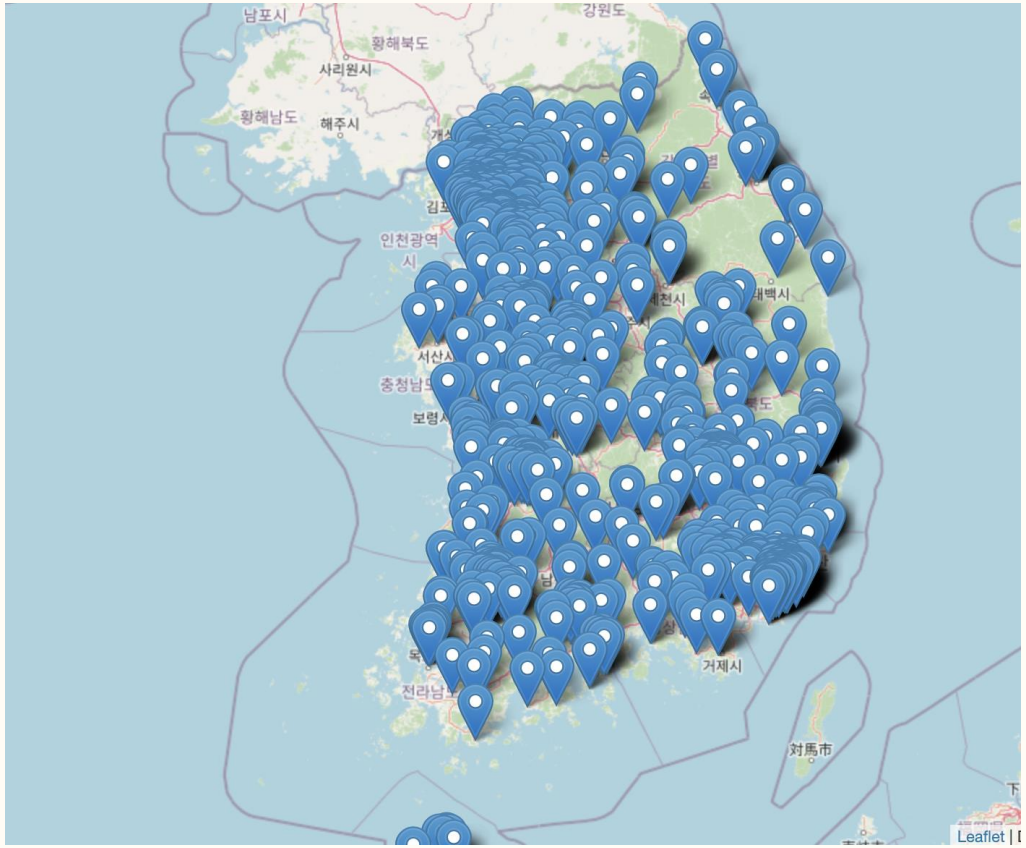
지도에서 노인/치매병원 위치 보기

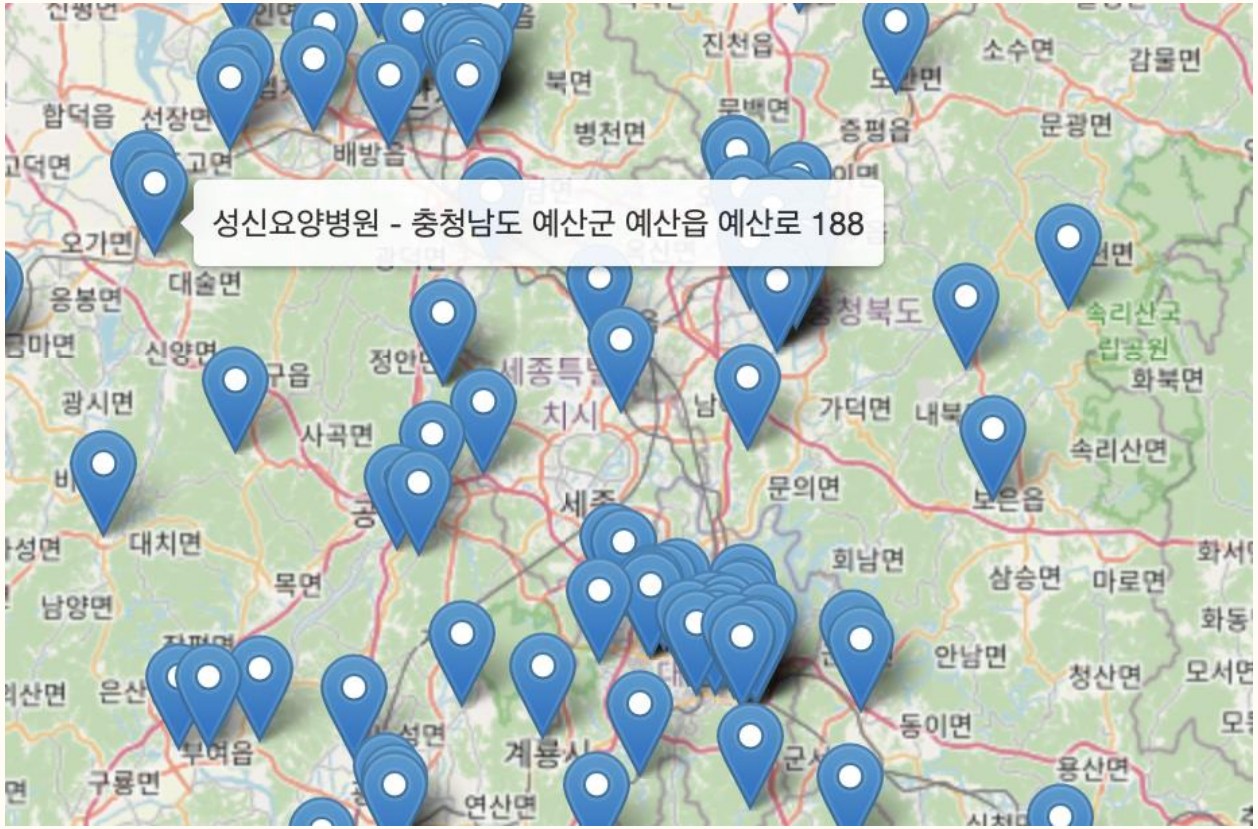
📌 Q4. 평균 기대수명이 점점 길어지면서 실버 의료 산업도 주목받고 있습니다.
여러분은 실버 의료 산업과 관련된 스타트업에 취업했습니다.

지도를 시각화하여 '노인/치매병원'이 주로 어디에 위치하고 있는지를 찾아보고자 합니다.
folium 을 통해 지도에 전국의 '노인/치매병원'을 표시해 주세요!

2주차 미션 - 4번

지도에서 노인/치매병원 위치 보기





NAVER CONNECT

모두의연구소

3주차 학습 내용

어떤 내용을 배우게 될까요?

건강검진 데이터 살펴보기: 메타 데이터

데이터를 모두 숫자로 표현

4	연령대 코드(5세 단위)	AGE_GROUP	• 기준년도에 수진자의 나이를 5세 단위로 그룹화(범주화)하여 구분한 코드 - 5세 단위 그룹화, 85세 이상은 85+로 그룹화 - 2002~2013년 까지	11	●

그룹	연령대	그룹	연령대
1	20~24세	8	55~59세
2	25~29세	9	60~64세
3	30~34세	10	65~69세
4	35~39세	11	70~74세
5	40~44세	12	75~79세
6	45~49세	13	80~84세
7	50~54세	14	85세+

• 2014년 이후

그룹	연령대	그룹	연령대
1	0~4세	10	45~49세
2	5~9세	11	50~54세
3	10~14세	12	55~59세
4	15~19세	13	60~64세
5	20~24세	14	65~69세
6	25~29세	15	70~74세
7	30~34세	16	75~79세
8	35~39세	17	80~84세
9	40~44세	18	85세+

5	시도코드	SIDO	• 해당 수진자 거주지의 시도코드 - 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음	N	26	●

코드명	시도명	코드명	시도명
11	서울특별시	42	강원도
26	부산광역시	43	충청북도
27	대구광역시	44	충청남도
28	인천광역시	45	전라북도
29	광주광역시	46	전라남도
30	대전광역시	47	경상북도
31	울산광역시	48	경상남도
36	세종특별자치시	49	제주특별자치도
41	경기도		

6	신장(5Cm단위)	HEIGHT	• 검진자의 키(5CM 단위) ❖ 예) 100~104CM -> 100CM	N/Cm	140	
7	체중(5Kg 단위)	WEIGHT	• 검진자의 몸무게(5KG 단위) ❖ 예) 25~29KG -> 25KG	N/Kg	45	

건강검진 데이터 살펴보기: 메타 데이터

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000000 entries, 0 to 999999
Data columns (total 34 columns):
 #   Column          Non-Null Count    Dtype
---  ------          --------------    -----
 0   기준년도          1000000 non-null  int64
 1   가입자일련번호       1000000 non-null  int64
 2   성별코드          1000000 non-null  int64
 3   연령대코드(5세단위)   1000000 non-null  int64
 4   시도코드          1000000 non-null  int64
 5   신장(5Cm단위)     999738 non-null   float64
 6   체중(5Kg 단위)    999732 non-null   float64
 7   허리둘레          999560 non-null   float64
 8   시력(좌)         999712 non-null   float64
 9   시력(우)         999721 non-null   float64
 10  청력(좌)         999758 non-null   float64
 11  청력(우)         999757 non-null   float64
 12  수축기혈압         999924 non-null   float64
 13  이완기혈압         999923 non-null   float64
 14  식전혈당(공복혈당)    999957 non-null   float64
 ...
 32  치석            398748 non-null   float64
 33  데이터공개일자       1000000 non-null  int64
dtypes: float64(27), int64(7)
memory usage: 259.4 MB
```

데이터를 숫자로 제공하는 이유?

텍스트 대비 효율적인 저장용량 관리

```
 32  치석            1000000 non-null  object
 33  데이터공개일자       1000000 non-null  object
dtypes: object(34)
memory usage: 259.4+ MB
```

일관된 값으로 정보의 품질 관리 가능

e.g. 서울, 서울특별시 등으로 같은 의미의 텍스트를 다르게 입력하면 관리가 어려움

건강검진 데이터 살펴보기: 도메인 지식

<table>
<tr>
<td>5</td>
<td>시도코드</td>
<td>SIDO</td>
<td>
• 해당 수진자 거주지의 시도코드
- 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음
<table>
<tr><th>코드명</th><th>시도명</th><th>코드명</th><th>시도명</th></tr>
<tr><td>11</td><td>서울특별시</td><td>42</td><td>강원도</td></tr>
<tr><td>26</td><td>부산광역시</td><td>43</td><td>충청북도</td></tr>
<tr><td>27</td><td>대구광역시</td><td>44</td><td>충청남도</td></tr>
<tr><td>28</td><td>인천광역시</td><td>45</td><td>전라북도</td></tr>
<tr><td>29</td><td>광주광역시</td><td>46</td><td>전라남도</td></tr>
<tr><td>30</td><td>대전광역시</td><td>47</td><td>경상북도</td></tr>
<tr><td>31</td><td>울산광역시</td><td>48</td><td>경상남도</td></tr>
<tr><td>36</td><td>세종특별자치시</td><td>49</td><td>제주특별자치도</td></tr>
<tr><td>41</td><td>경기도</td><td></td><td></td></tr>
</table>
</td>
<td>N</td>
<td>26</td>
<td>●</td>
</tr>
</table>

건강검진 데이터 살펴보기: 도메인 지식

김부스터 씨,
데이터 살펴보고
'시도코드'를
문자로 변경 좀 해주세요

건강검진 데이터 살펴보기: 도메인 지식

부스트코스가
벌써 2주차가 지났는데
그 정도는 껌이죠!

PDF 를 보고 제가
해드리죠

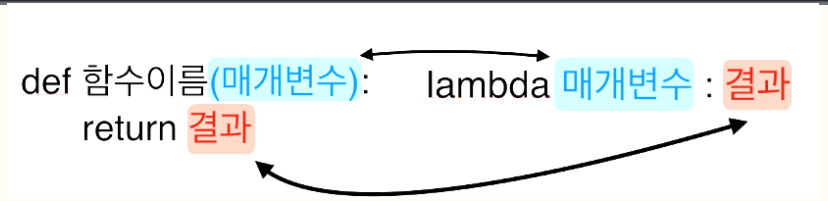
건강검진 데이터 살펴보기: 도메인 지식

- 해당 수진자 거주지의 시도코드
  - 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음

코드명	시도명	코드명	시도명
11	서울특별시	42	강원도
26	부산광역시	43	충청북도
27	대구광역시	44	충청남도
28	인천광역시	45	전라북도
29	광주광역시	46	전라남도
30	대전광역시	47	경상북도
31	울산광역시	48	경상남도
36	세종특별자치시	49	제주특별자치도
41	경기도		

건강검진 데이터 살펴보기: 도메인 지식

```python
code = {11: '서울특별시', 26: '부산광역시', 27: '대구광역시', 28: '인천광역시', 29: '광주광역시',
        30: '대전광역시', 31: '울산광역시', 36: '세종특별자치시', 41: '경기도', 42: '강원도',
        43: '충청북도', 44: '충청남도', 45: '전라북도', 46: '전라남도', 47: '경상북도',
        48: '경상남도', 49: '제주특별자치도'}

df['시도코드'] = df['시도코드'].apply(lambda x: code[x])
```



람다함수는 결과부분된 부분을 return키워드 없이 자동으로 return해줍니다.

건강검진 데이터 살펴보기: 도메인 지식

```
KeyError                                  Traceback (most recent call last)
Cell In[26], line 8
      1 df = pd.read_csv("https://raw.githubusercontent.com/corazzon/boostcourse-ds-510/master/data/NHIS_OPEN_GJ_2017.CSV.zip",
      2                  encoding="cp949")
      4 code = {11: '서울특별시', 26: '부산광역시', 27: '대구광역시', 28: '인천광역시', 29: '광주광역시', 30: '대전광역시',
      5         31: '울산광역시', 36: '세종특별자치시', 41: '경기도', 42: '강원도', 43: '충청북도', 44: '충청남도', 45: '전라북도',
      6         46: '전라남도', 47: '경상북도', 48: '경상남도', 49: '제주특별자치도'}
----> 8 df['시도코드'] = df['시도코드'].apply(lambda x: code[x])

Cell In[26], line 8, in <lambda>(x)
      1 df = pd.read_csv("https://raw.githubusercontent.com/corazzon/boostcourse-ds-510/master/data/NHIS_OPEN_GJ_2017.CSV.zip",
      2                  encoding="cp949")
      4 code = {11: '서울특별시', 26: '부산광역시', 27: '대구광역시', 28: '인천광역시', 29: '광주광역시', 30: '대전광역시',
      5         31: '울산광역시', 36: '세종특별자치시', 41: '경기도', 42: '강원도', 43: '충청북도', 44: '충청남도', 45: '전라북도',
      6         46: '전라남도', 47: '경상북도', 48: '경상남도', 49: '제주특별자치도'}
----> 8 df['시도코드'] = df['시도코드'].apply(lambda x: code[x])

KeyError: 50
```

```
   1154     else:
   1155         values = obj.astype(object)._values
-> 1156         mapped = lib.map_infer(
   1157             values,
   1158             f,
   1159             convert=self.convert_dtype,
   1160         )
   1162 if len(mapped) and isinstance(mapped[0], ABCSeries):
   1163     # GH#43986 Need to do list(mapped) in order to get treated as nested
   1164     #  See also GH#25959 regarding EA support
   1165     return obj._constructor_expanddim(list(mapped), index=obj.index)

File ~\miniconda3\lib\site-packages\pandas\_libs\lib.pyx:2918, in pandas._libs.lib.map_infer()
```


```
Cell In[26], line 8, in <lambda>(x)
      1 df = pd.read_csv("https://raw.githubusercontent.com/corazzon/boostcourse-ds-510/master/data/NHIS_OPEN_GJ_2017.CSV.zip",
      2                  encoding="cp949")
      4 code = {11: '서울특별시', 26: '부산광역시', 27: '대구광역시', 28: '인천광역시', 29: '광주광역시', 30: '대전광역시',
      5         31: '울산광역시', 36: '세종특별자치시', 41: '경기도', 42: '강원도', 43: '충청북도', 44: '충청남도', 45: '전라북도',
      6         46: '전라남도', 47: '경상북도', 48: '경상남도', 49: '제주특별자치도'}
----> 8 df['시도코드'] = df['시도코드'].apply(lambda x: code[x])

KeyError: 50
```

건강검진 데이터 살펴보기: 도메인 지식

KEY 50?????

건강검진 데이터 살펴보기: 도메인 지식

5	시도코드	SIDO

- 해당 수진자 거주지의 시도코드
  - 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음

코드명	시도명	코드명	시도명
11	서울특별시	42	강원도
26	부산광역시	43	충청북도
27	대구광역시	44	충청남도
28	인천광역시	45	전라북도
29	광주광역시	46	전라남도
30	대전광역시	47	경상북도
31	울산광역시	48	경상남도
36	세종특별자치시	49	제주특별자치도
41	경기도		

50 은 없는데?
서울 11? 부산 26?

건강검진 데이터 살펴보기: 도메인 지식

https://www.code.go.kr/stdcode/regCodeL.do

건강검진 데이터 살펴보기: 도메인 지식

5	시도코드	SIDO	• 해당 수진자 거주지의 시도코드 - 2012년부터 세종특별자치시가 신규로 편입됨에 따라, 2011년까지의 데이터에는 해당 항목이 존재하지 않음

코드명	시도명	코드명	시도명
11	서울특별시	42	강원도
26	부산광역시	43	충청북도
27	대구광역시	44	충청남도
28	인천광역시	45	전라북도
29	광주광역시	46	전라남도
30	대전광역시	47	경상북도
31	울산광역시	48	경상남도
36	세종특별자치시	49	제주특별자치도
41	경기도		

```
df['시도코드'].value_counts().sort_index()
```

```
11    177346
26     68642
27     47487
28     58682
29     28519
30     30237
31     25327
36      4720
41    243865
42     30689
43     34061
44     42822
45     36704
46     37972
47     54132
48     67735
50     11060
Name: 시도코드, dtype: int64
```

건강검진 데이터 살펴보기: 도메인 지식

```
code = {11: '서울특별시', 26: '부산광역시', 27: '대구광역시', 28: '인천광역시', 29: '광주광역시',
        30: '대전광역시', 31: '울산광역시', 36: '세종특별자치시', 41: '경기도', 42: '강원도',
        43: '충청북도', 44: '충청남도', 45: '전라북도', 46: '전라남도', 47: '경상북도',
        48: '경상남도', 50: '제주특별자치도'}

df['시도코드'] = df['시도코드'].apply(lambda x: code[x])
```

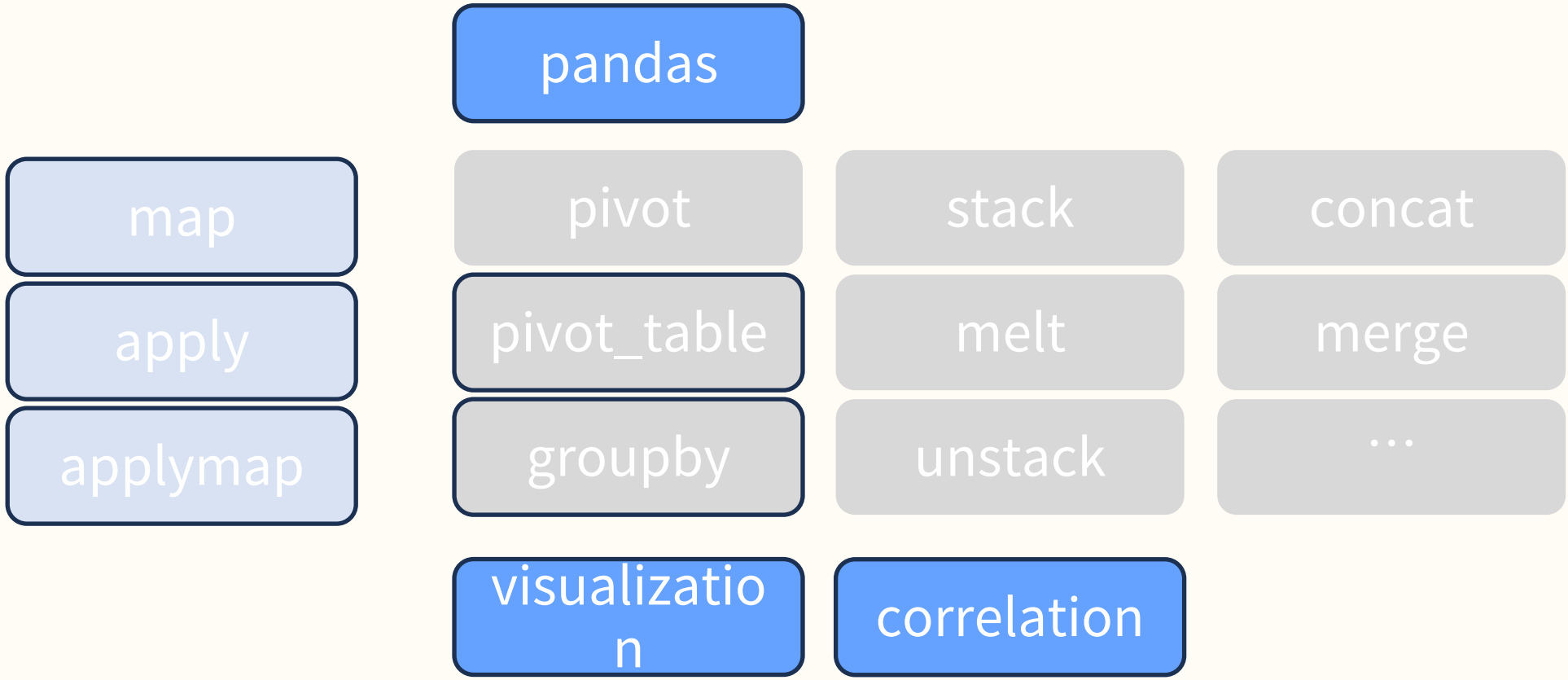
건강검진 데이터 살펴보기: 도메인 지식

```
df['시도코드'].value_counts()
```

```
경기도          243865
서울특별시       177346
부산광역시        68642
경상남도         67735
인천광역시        58682
경상북도         54132
대구광역시        47487
충청남도         42822
전라남도         37972
전라북도         36704
충청북도         34061
강원도          30689
대전광역시        30237
광주광역시        28519
울산광역시        25327
제주특별자치도      11060
세종특별자치시       4720
Name: 시도코드, dtype: int64
```

건강검진 데이터 살펴보기: 도메인 지식

건강검진 데이터 다뤄보기: 이번 주에 배울 내용들



건강검진 데이터 다뤄보기: map, apply

	Series	DataFrame
map	**O**	**X**
apply	**O**	**O**

map: Series 에 사용되며 Series 의 각 요소에 특정 함수 또는 mapping 을 적용

apply: DataFrame 과 Series 모두에 사용되며 행이나 열에 특정 함수를 적용

건강검진 데이터 다뤄보기: map

map: Series 에 사용되며 Series 의 각 요소에 특정 함수 또는 mapping 을 적용

```
# 1. Using map(): Convert the gender code (성별코드)
# from numbers to a more readable format
gender_mapping = {1: 'Male', 2: 'Female'}
df['성별코드'] = df['성별코드'].map(gender_mapping)

# Displaying the updated gender column
df[['성별코드']].head()
```

	성별코드
0	Male
1	Female
2	Male
3	Female
4	Male

건강검진 데이터 다뤄보기: apply

apply: DataFrame 과 Series 모두에 사용되며 행이나 열에 특정 함수를 적용

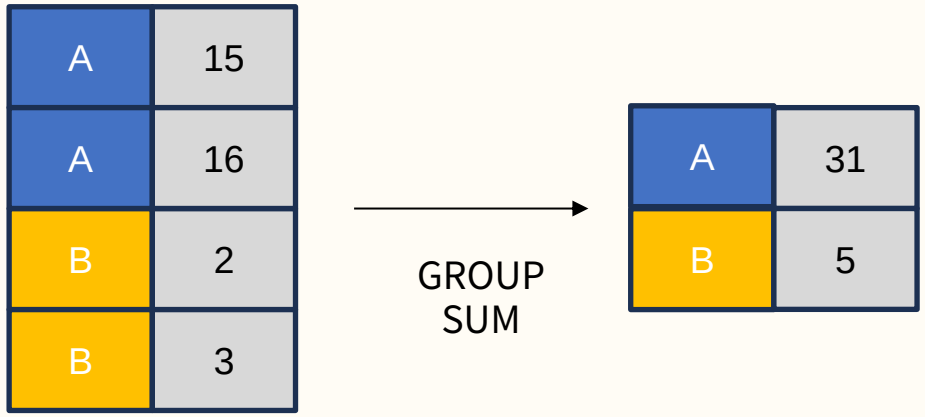
```python
# 2. Using apply(): Compute the BMI for each individual
def compute_bmi(row):
    # Convert height from cm to meters
    height_m = row['신장(5Cm단위)'] / 100
    weight = row['체중(5Kg 단위)']
    bmi = weight / (height_m ** 2)
    return bmi

df['Approx.BMI'] = df.apply(compute_bmi, axis=1)

# Displaying the first few rows with the new BMI column
df[['신장(5Cm단위)', '체중(5Kg 단위)', 'Approx.BMI']].head()
```

	신장(5Cm단위)	체중(5Kg 단위)	Approx.BMI
0	170.0	65.0	22.491349
1	150.0	45.0	20.000000
2	175.0	75.0	24.489796
3	155.0	55.0	22.892820
4	175.0	75.0	24.489796

건강검진 데이터 다뤄보기: groupby

groupby

group data based on some criteria
and then apply a function to each group of data.



	기능	사용법
agg	요약된 통계정보	grouped.agg({‘B’: mean}
transform	정보의 변환	grouped.transform()
filter	정보의 필터링	grouped.filter(filter_func)

건강검진 데이터 다뤄보기: groupby & agg

```
agg_df = df.groupby('성별코드').agg({
    '신장(5Cm단위)': 'mean',
    '체중(5Kg 단위)': 'mean',
})
agg_df.head()
```

성별코드	신장(5Cm단위)	체중(5Kg 단위)
1	168.516441	70.211016
2	155.078899	55.701656

인사 데이터 다뤄보기: groupby & transform

	Name	Company	Department	Salary
0	Alice	Google	HR	70000
1	Bob	Google	Engineering	120000
2	Charlie	Apple	HR	80000
3	David	Apple	Engineering	110000
4	Edward	Amazon	HR	90000
5	Fiona	Amazon	Engineering	95000
6	George	Amazon	Engineering	105000

```python
# 각 회사별로 그룹화
grouped = df.groupby('Company')

# 각 회사의 평균 연봉 계산
average_salary = grouped['Salary'].transform('mean')

# 각 직원의 연봉이 해당 회사의 평균 연봉보다 높은지 여부를 나타내는 새로운 열 추가
df['Above Average'] = df['Salary'] > average_salary
df
```

	Name	Company	Department	Salary	Above Average
0	Alice	Google	HR	70000	False
1	Bob	Google	Engineering	120000	True
2	Charlie	Apple	HR	80000	False
3	David	Apple	Engineering	110000	True
4	Edward	Amazon	HR	90000	False
5	Fiona	Amazon	Engineering	95000	False
6	George	Amazon	Engineering	105000	True

인사 데이터 다뤄보기: groupby & filter

	Name	Company	Department	Salary
0	Alice	Google	HR	70000
1	Bob	Google	Engineering	120000
2	Charlie	Apple	HR	80000
3	David	Apple	Engineering	110000
4	Edward	Amazon	HR	90000
5	Fiona	Amazon	Engineering	95000
6	George	Amazon	Engineering	105000

```python
# 회사별로 그룹화
grouped = df.groupby('Company')

# 직원 수가 3명 이상인 회사만 필터링
filtered_df = grouped.filter(lambda x: len(x) >= 3)
filtered_df
```

	Name	Company	Department	Salary
4	Edward	Amazon	HR	90000
5	Fiona	Amazon	Engineering	95000
6	George	Amazon	Engineering	105000

건강검진 데이터 다뤄보기: Pivot

Pivot

Pivot 은 연산이 아닌 형태 변환(Reshape) 용

```
df.pivot(index='foo',
         columns='bar',
         values='baz')
```

df

	foo	bar	baz	zoo
0	one	A	1	x
1	one	B	2	y
2	one	C	3	z
3	two	A	4	q
4	two	B	5	w
5	two	C	6	t

bar	A	B	C
foo			
one	1	2	3
two	4	5	6

https://pandas.pydata.org/docs/user_guide/reshaping.html#pivot-and-pivot-table

건강검진 인사 데이터 다뤄보기: pivot_table

	Name	Company	Department	Salary
0	Alice	Google	HR	70000
1	Bob	Google	Engineering	120000
2	Charlie	Apple	HR	80000
3	David	Apple	Engineering	110000
4	Edward	Amazon	HR	90000
5	Fiona	Amazon	Engineering	95000
6	George	Amazon	Engineering	105000

```
df = pd.DataFrame(data)

# 회사별 평균 급여 계산
pivot_table = df.pivot_table(values='Salary', index='Company', aggfunc='mean')
pivot_table
```

RESULT

Company	Salary
Amazon	96666.666667
Apple	95000.000000
Google	95000.000000

데이터 시각화의 목적

많은 양의 데이터를 한 눈에!

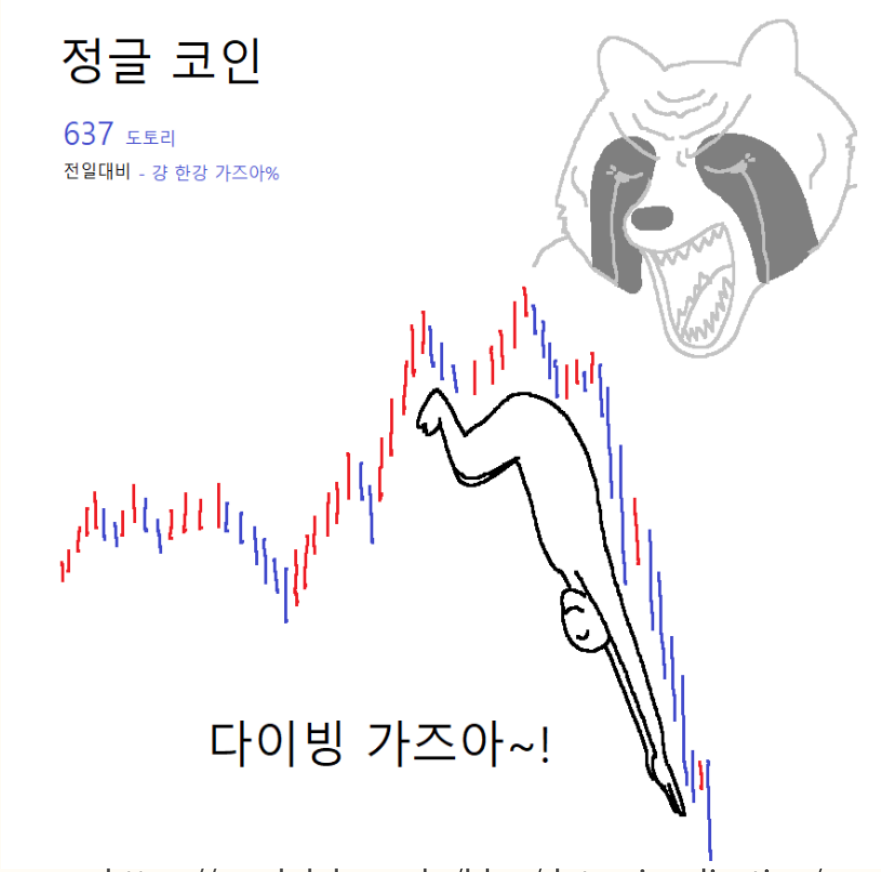
데이터에 대한 인사이트

요약 통계보다 정확한 분석 가능

의미 있는 커뮤니케이션

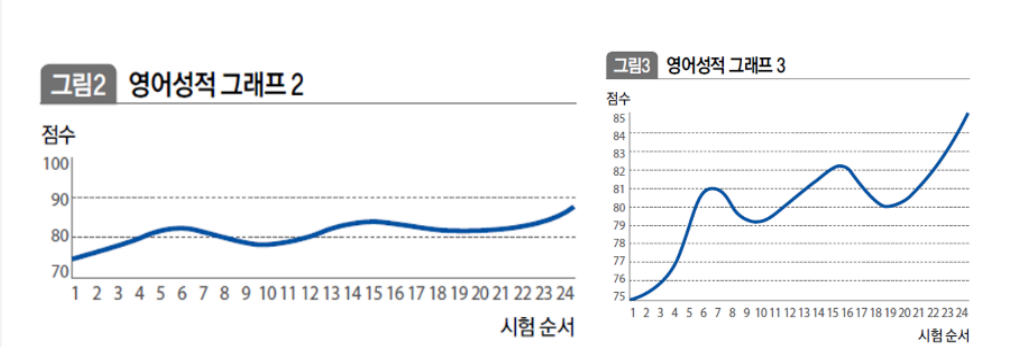
https://newsjel.ly/archives/newsjelly-report/visualization-report/8136

데이터 시각화 주의사항



https://modulabs.co.kr/blog/data-visualization/

데이터의 왜곡 주의



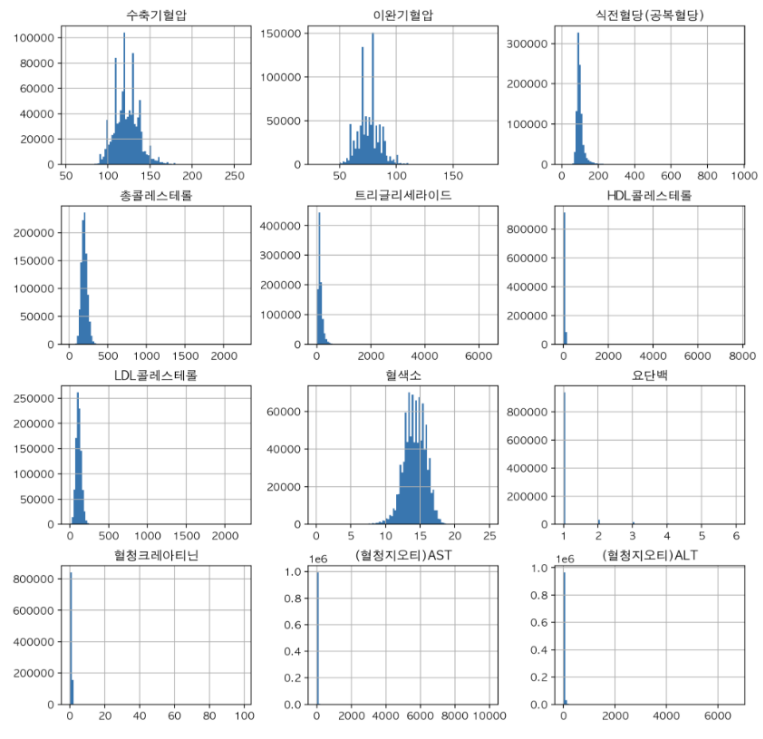
https://m.blog.naver.com/businessinsight/221918586252

시각화에만 의존한 의사 판단 주의

대용량 데이터를 시각화 할 때는?

- 100만개가 넘는 데이터를 시각화할 때는 되도록이면 groupby 혹은 pivot_table로 연산을 하고 시각화를 하는 것을 권장합니다.
- 100만개가 넘는 데이터를 seaborn과 같은 고급 통계 연산을 하는 그래프를 사용하게 되면 많이 느릴 수 있습니다.

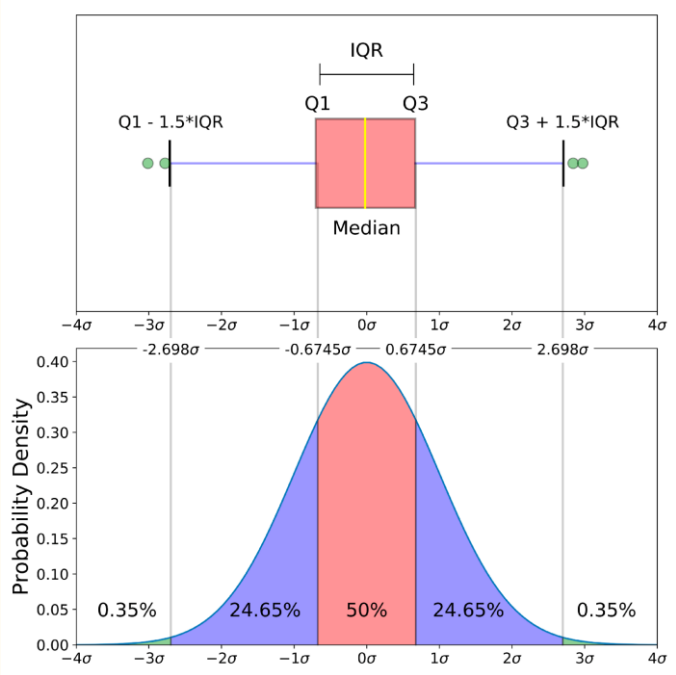
히스토그램



- **히스토그램(histogram)**은 표로 되어 있는 **도수 분포를 정보 그림으로 나타낸 것**이다.
- **히스토그램: 도수분포표를 그래프로 나타낸 것**
- 보통 히스토그램에서는 가로축이 계급, 세로축이 도수 때때로 반대로 그리기도 한다.
- 계급은 보통 변수의 구간이고, 서로 겹치지 않는다.
- 그림에서 계급(막대기)끼리는 서로 붙어 있어야 한다.
- 히스토그램은 일반 막대그래프와는 다르다. 막대그래프는 계급 즉 가로를 생각하지 않고 세로의 높이로만 나타내지만 히스토그램은 가로와 세로를 함께 생각해야 한다.

https://ko.wikipedia.org/wiki/히스토그램

박스플롯 상자수염그림

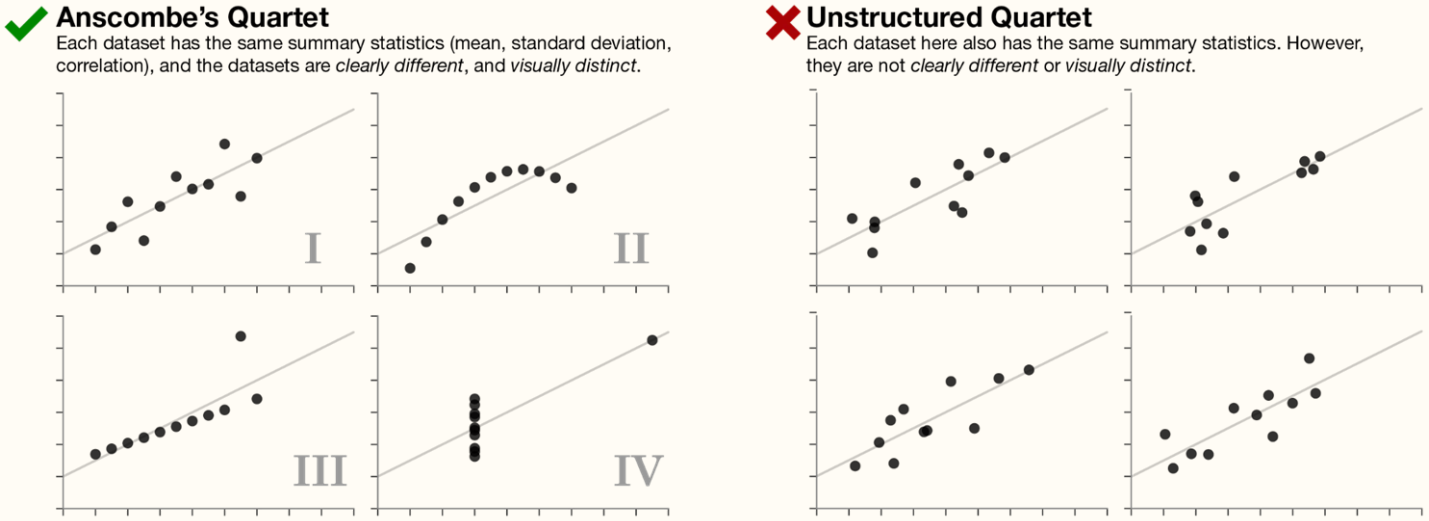


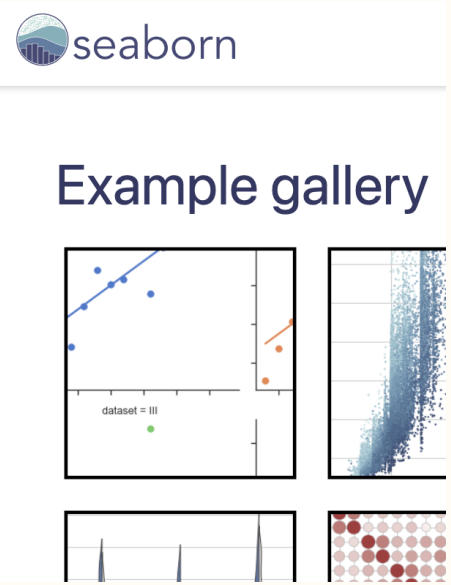
https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

- 백분위 수 : 데이터를 백등분 한 것
- 사분위 수 : 데이터를 4등분 한 것
- 중위수 : 데이터의 정 가운데 순위에 해당하는 값.(관측치의 절반은 크거나 같고 나머지 절반은 작거나 같다.)
- 제 3사분위 수 (Q3) : 중앙값 기준으로 상위 50% 중의 중앙값, 전체 데이터 중 상위 25%에 해당하는 값
- 제 1사분위 수 (Q1) : 중앙값 기준으로 하위 50% 중의 중앙값, 전체 데이터 중 하위 25%에 해당하는 값
- 사분위 범위 수(IQR) : 데이터의 중간 50% (Q3 - Q1)

https://ko.wikipedia.org/wiki/상자_수염_그림

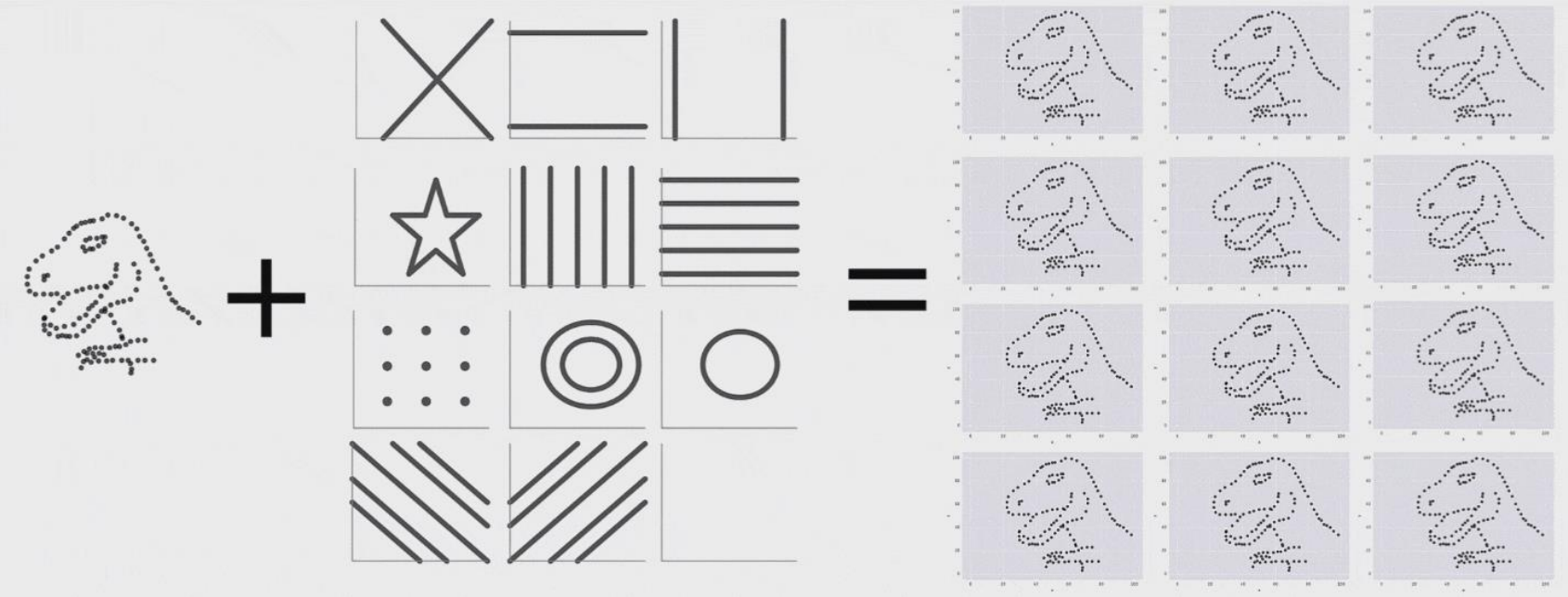
데이터 시각화의 중요성 - Anscombe's Quartet





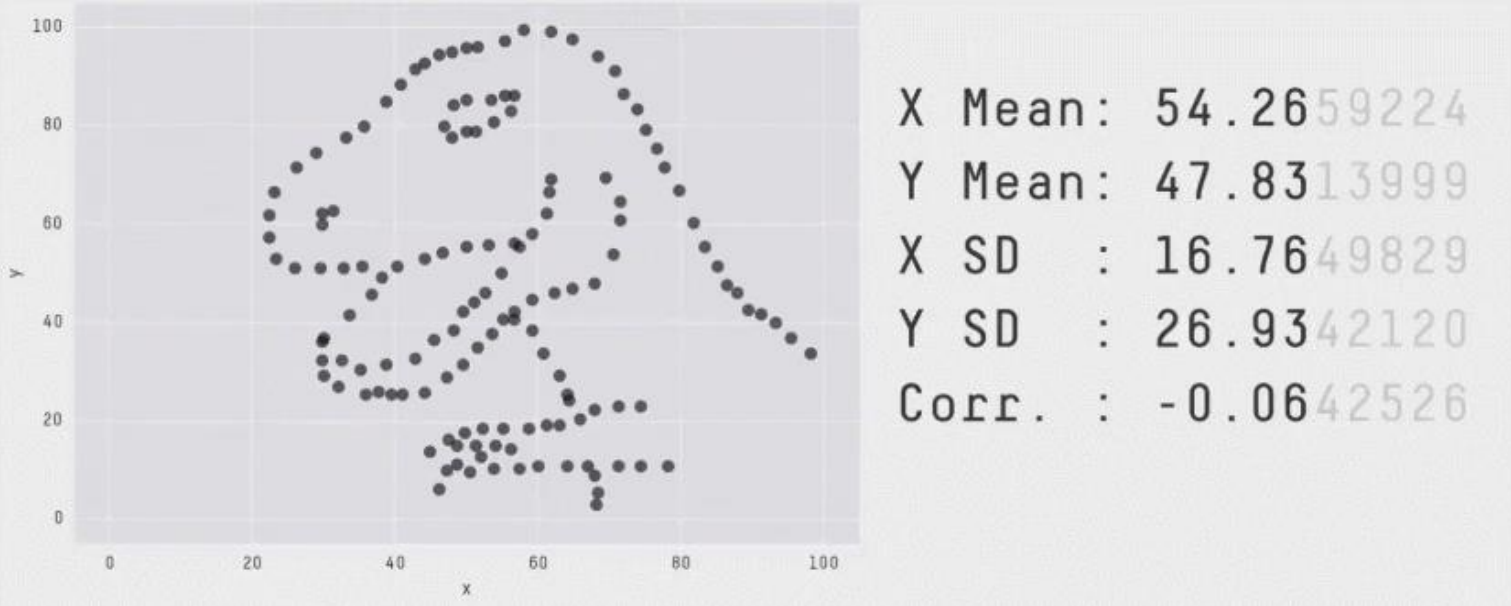
https://www.autodesk.com/research/publications/same-stats-different-graphs

같은 통계량이지만 다른 시각화



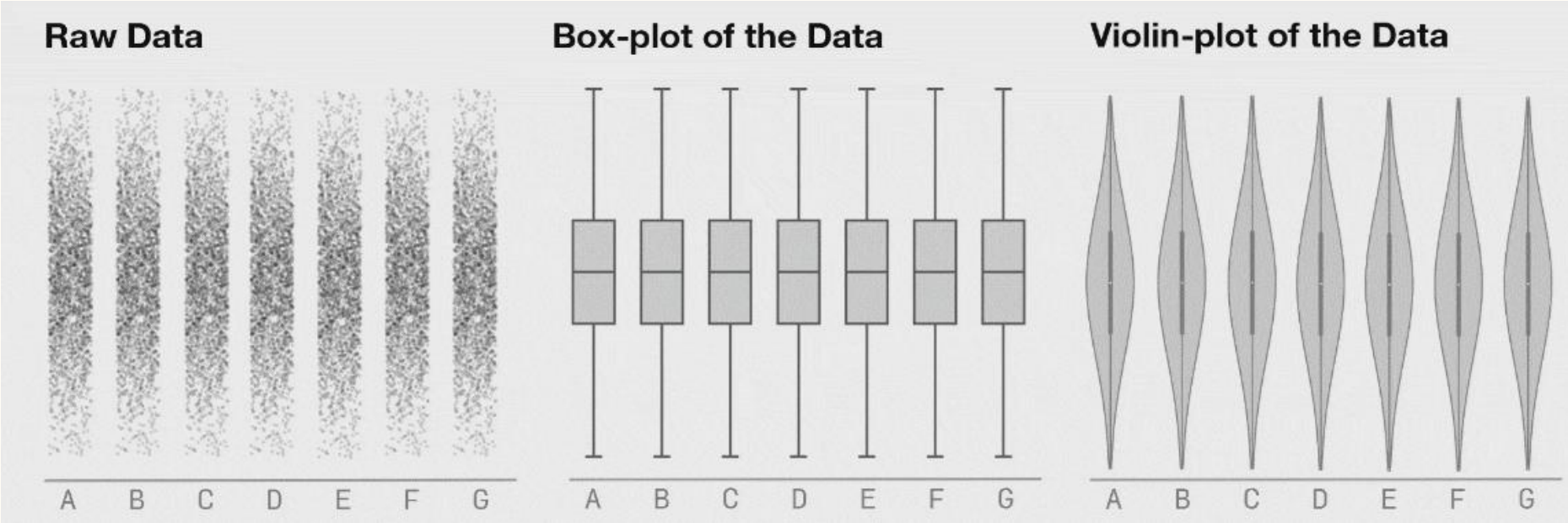
https://www.autodesk.com/research/publications/same-stats-different-graphs

같은 통계량이지만 다른 시각화



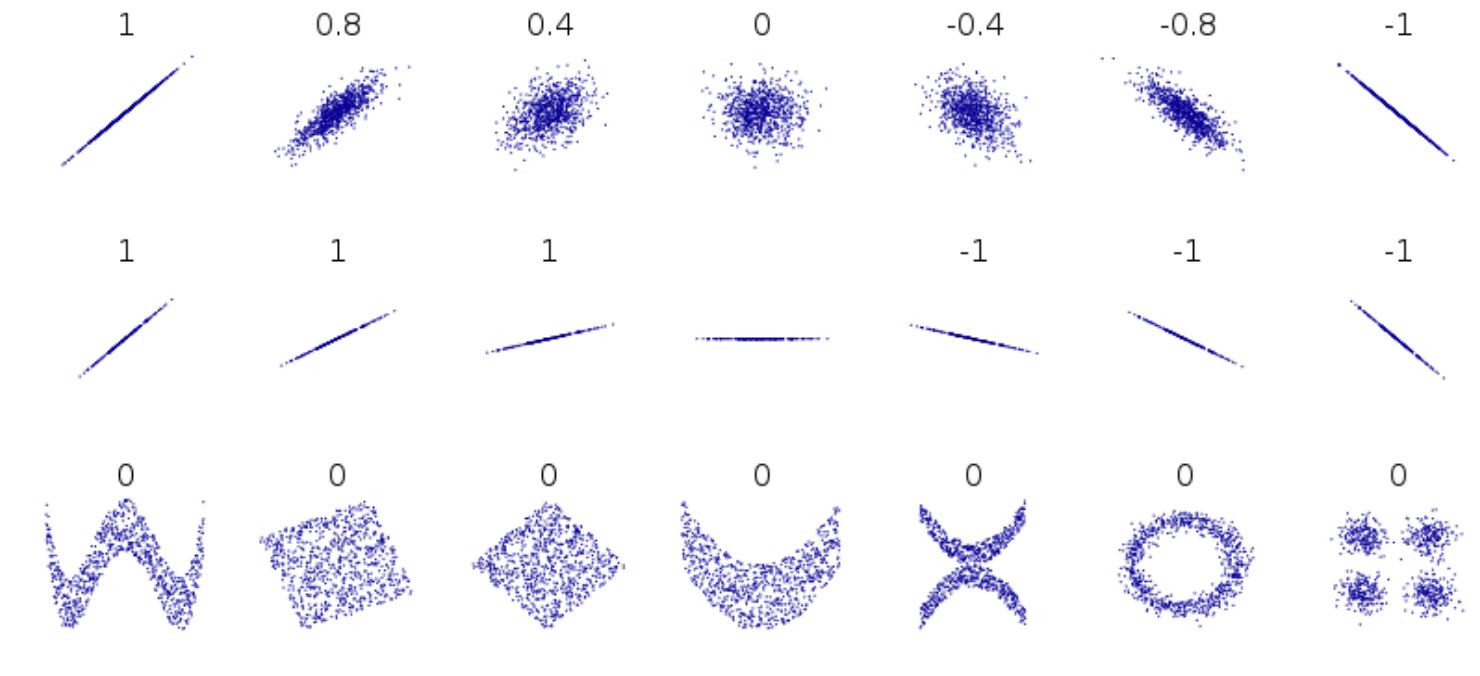
https://www.autodesk.com/research/publications/same-stats-different-graphs

box-plot 의 단점을 보완하여 scatterplot, violin-plot 으로 표현



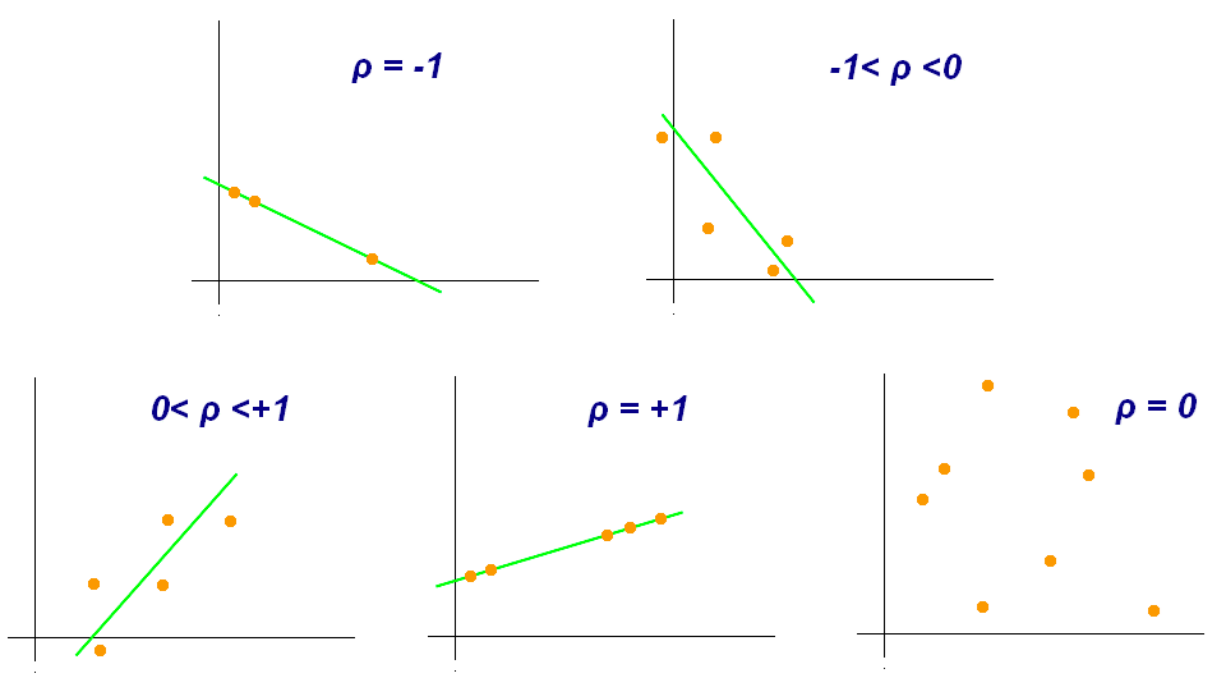
https://www.autodesk.com/research/publications/same-stats-different-graphs

상관 분석



https://ko.wikipedia.org/wiki/상관_분석#피어슨_상관_계수

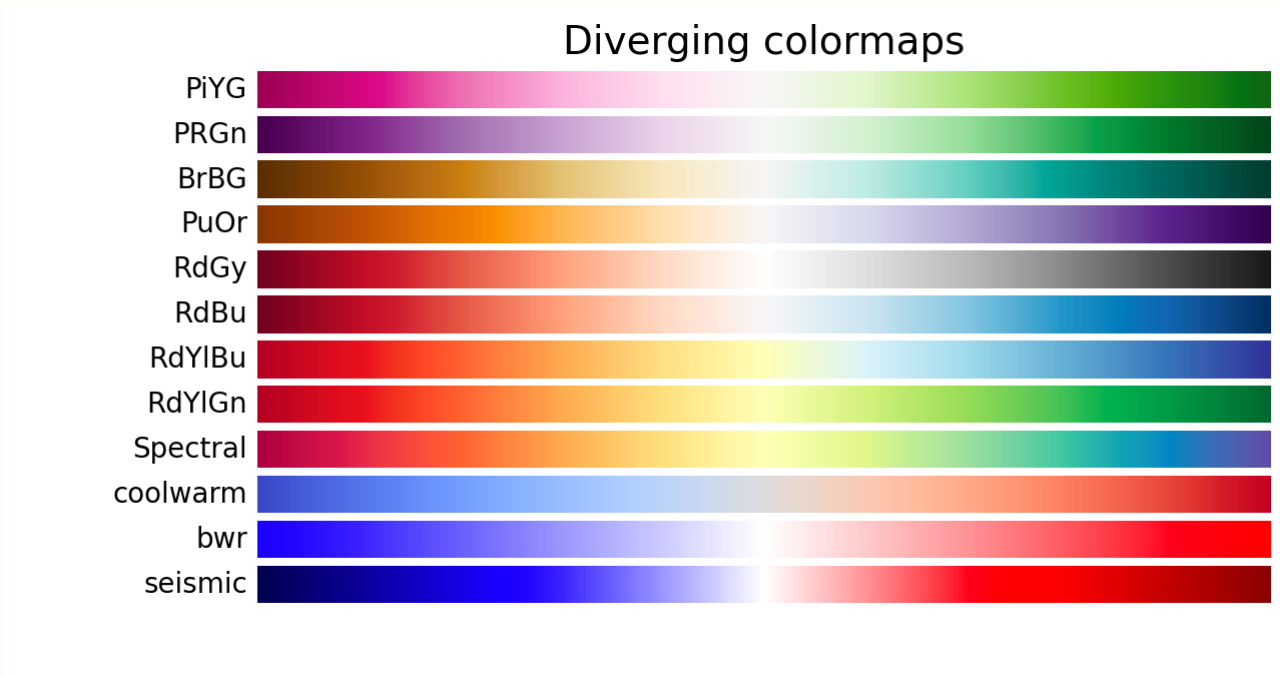
상관 분석



https://ko.wikipedia.org/wiki/상관_분석#피어슨_상관_계수

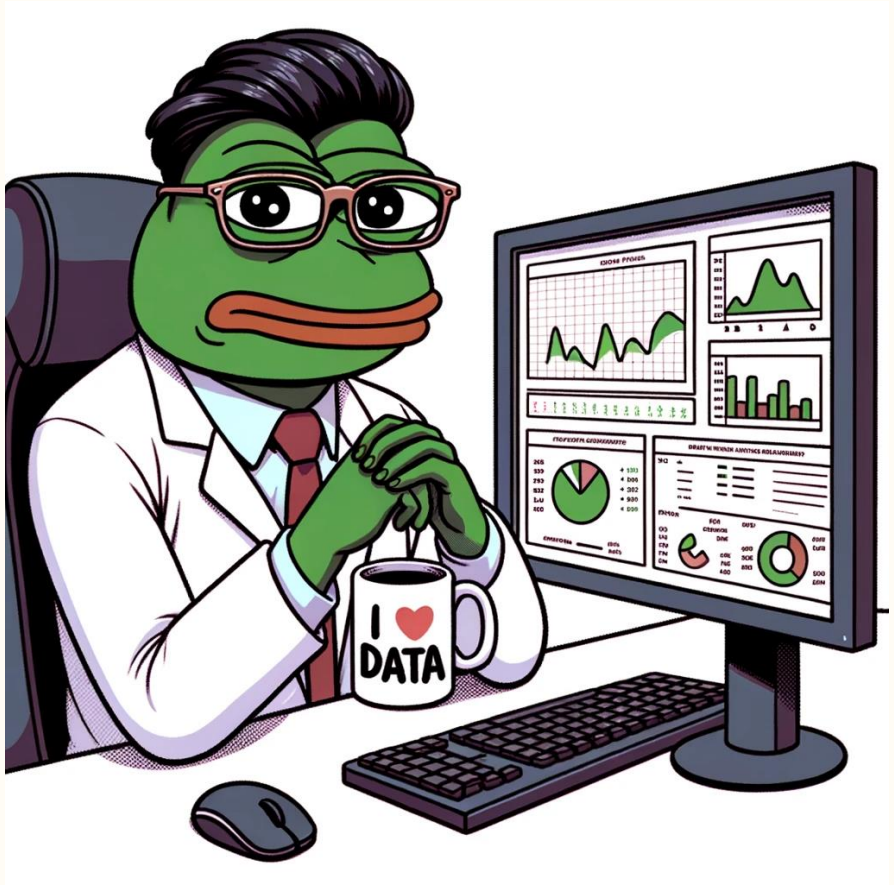
- r이 -1.0과 -0.7 사이이면, 강한 음적 선형관계
- r이 -0.7과 -0.3 사이이면, 뚜렷한 음적 선형관계
- r이 -0.3과 -0.1 사이이면, 약한 음적 선형관계
- r이 -0.1과 +0.1 사이이면, 거의 무시될 수 있는 선형관계
- r이 +0.1과 +0.3 사이이면, 약한 양적 선형관계
- r이 +0.3과 +0.7 사이이면, 뚜렷한 양적 선형관계
- r이 +0.7과 +1.0 사이이면, 강한 양적 선형관계

Matplotlib Diverging maps



https://matplotlib.org/stable/tutorials/colors/colormaps.html

알아두면 쓸데있는 신비한 프로그래밍: AI 와 함께 일해보자! (feat. ChatGPT)



역할 부여

↓

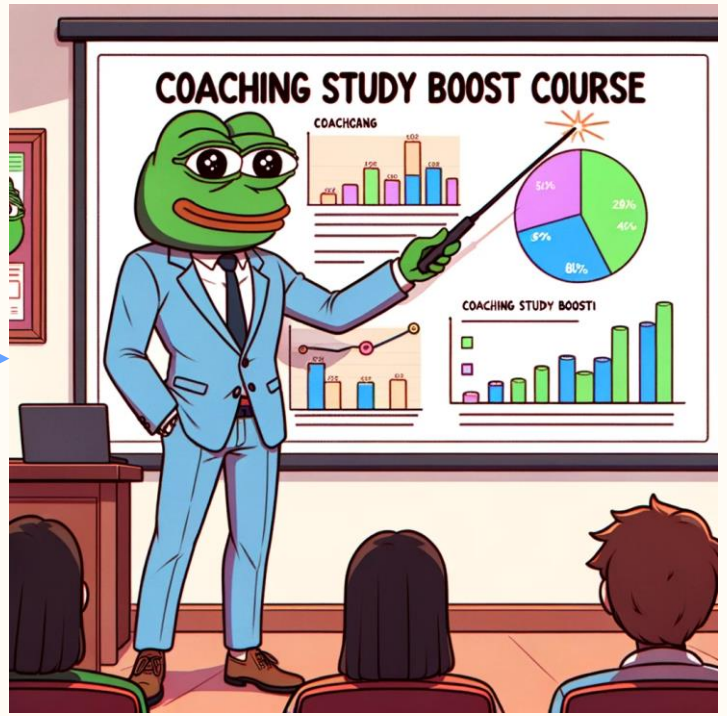
데이터 제공 — DATA

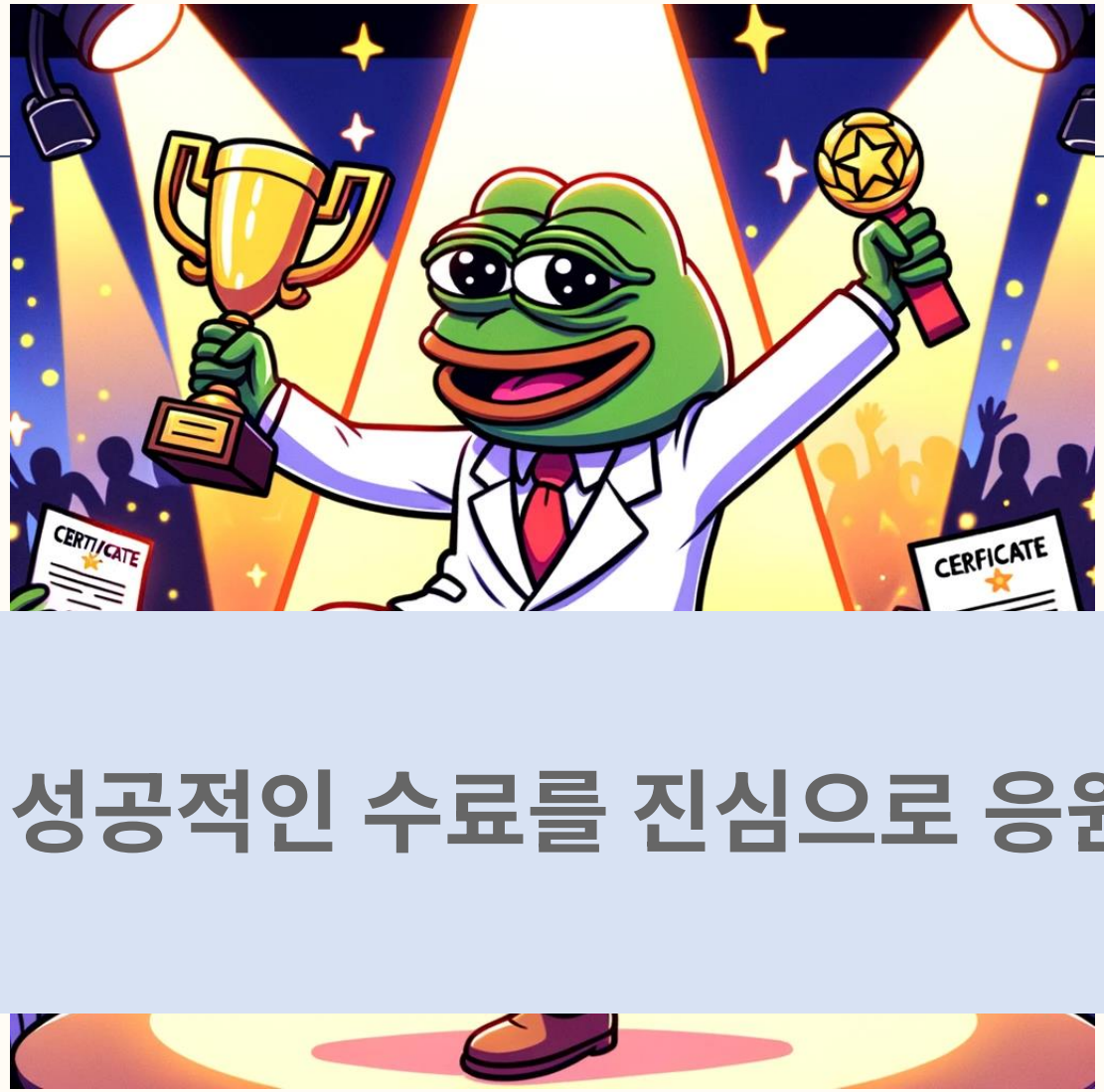
↓

업무 부여

알아두면 쓸데있는 신비한 프로그래밍: AI 와 함께 일해보자! (feat. ChatGPT)

역할 부여





여러분의 성공적인 수료를 진심으로 응원 합니다!

미션 출제 의도

3주차

3주차 미션의 출제 의도와 문제 소개

3주차 미션 출제 의도

1번	연령대별 허리둘레에 대한 기술통계를 구하기
2번	“음주여부", "흡연상태", "연령대코드(5세단위)", "성별코드" 상관계수를 구와 시각화 해보기
3번	흡연하는 사람과 음주하는 사람들의 수의 차이 알기
4번	체중이 120Kg 이상인 사람의 "총콜레스테롤", "감마지티피" 값을 음주여부에 따라 산점도로 시각화 해보기
5번	연령대별로 시력은 얼마나 차이날지 알아보기

3주차 미션 - 1번

기술 통계

📌 Q1.

연령대별 허리둘레에 대한 기술통계를 구하려고 합니다.

다음 제공되는 딕셔너리를 통해 연령대코드(5세단위)를 "연령대"로 만들고 아래와 같은 기술통계값을 구해주세요!

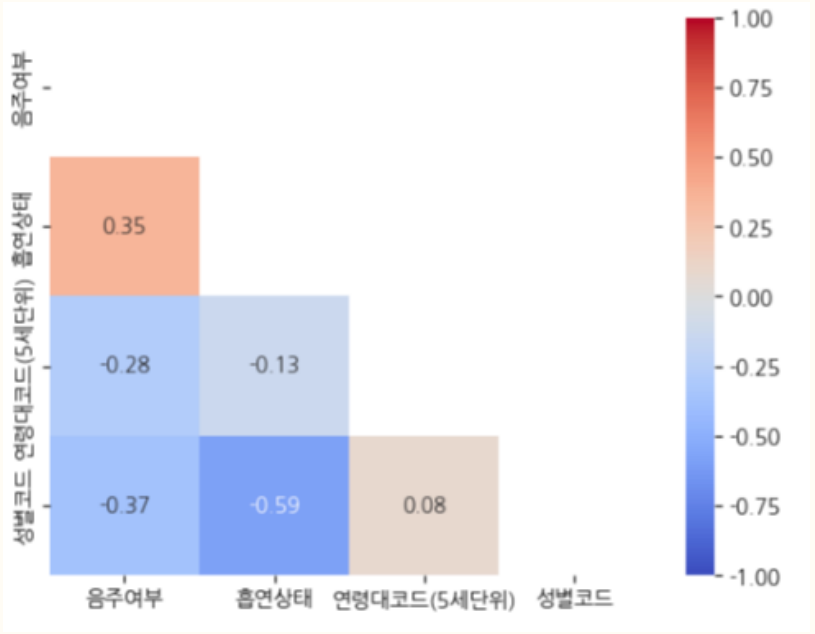
연령대	count	mean	std	min	25%	50%	75%	max
20~24세	23244	75.1522	12.2518	47.5	67.5	73.4	81	999
25~29세	64898	77.7048	16.7357	48	69	76.5	84.2	999
30~34세	77517	81.0893	22.9881	49	72	80.1	88	999
35~39세	84621	82.094	14.5221	9.2	75	82	89	999
40~44세	130912	80.4883	10.8031	42.1	73	80	87	999
45~49세	118357	80.8224	9.52162	40	74	81	87	137
50~54세	129833	81.0628	9.09544	6.5	75	81	87	142
55~59세	112175	81.7999	8.7304	32	76	82	87.5	139
60~64세	106491	82.7228	8.59618	0	77	83	88	137
65~69세	53624	83.5885	8.44354	50	78	83.5	89	129
70~74세	51586	84.0634	8.53964	51	78	84	90	129.8
75~79세	25972	84.2001	8.77231	50	78	84	90	122
80~84세	16205	83.7514	9.04109	38	78	84	90	120
85세+	4125	81.7367	17.326	34	75	81.5	88	999

3주차 미션 - 2번

상관계수와 시각화

📌 Q2.

"음주여부", "흡연상태", "연령대코드(5세단위)", "성별코드"에 대한 상관계수를 구하고 시각화 해주세요.



3주차 미션 - 3번

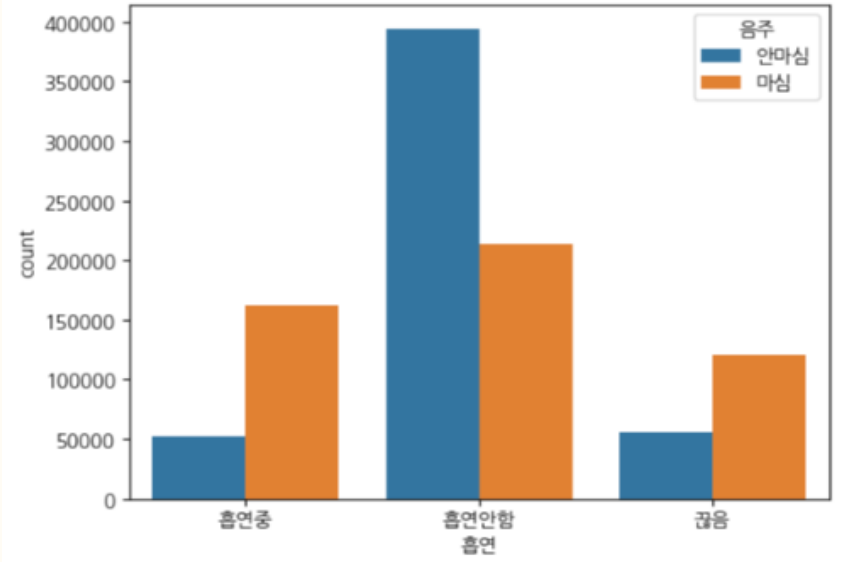
데이터 비교

📌 Q3.

흡연하는 사람과 음주하는 사람들의 수는 얼마나 차이가 있을까요?

결과 예시

음주	끊음	흡연안함	흡연중
마심	120779	213743	162166
안마심	55334	394503	52845

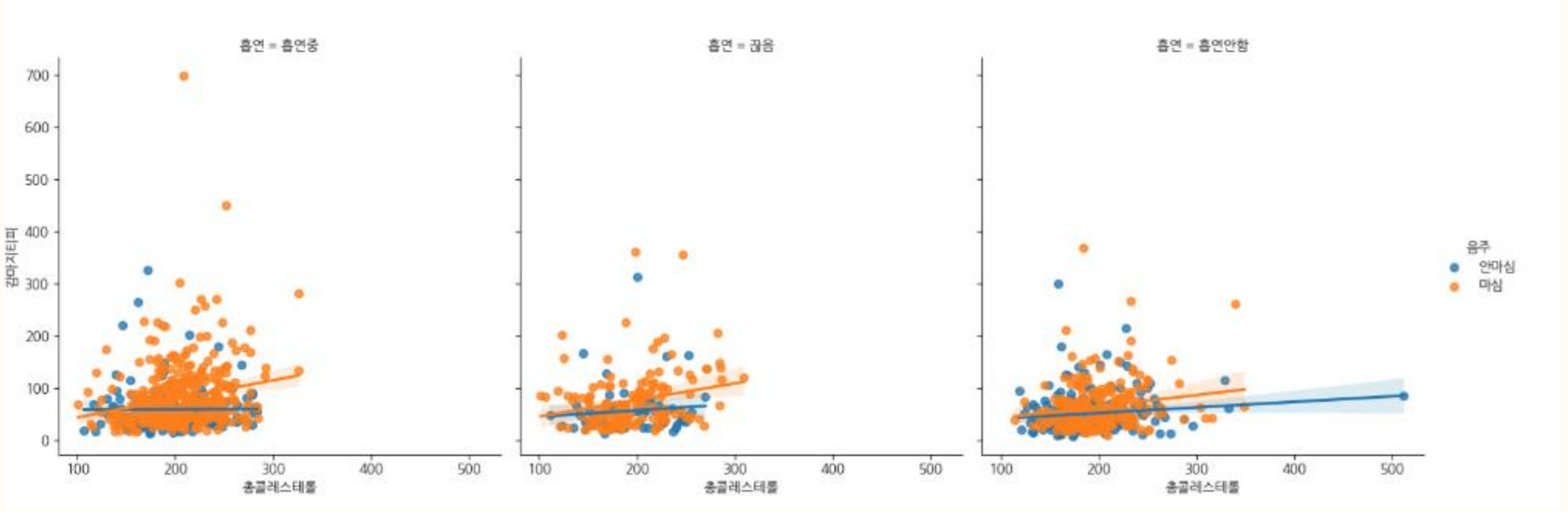


3주차 미션 - 4번

조건 있는 산점도 그리기

📌 Q4.

체중이 120Kg 이상인 데이터를 찾아 "총콜레스테롤", "감마지티피" 값을 음주여부에 따라 산점도로 시각화해주세요!

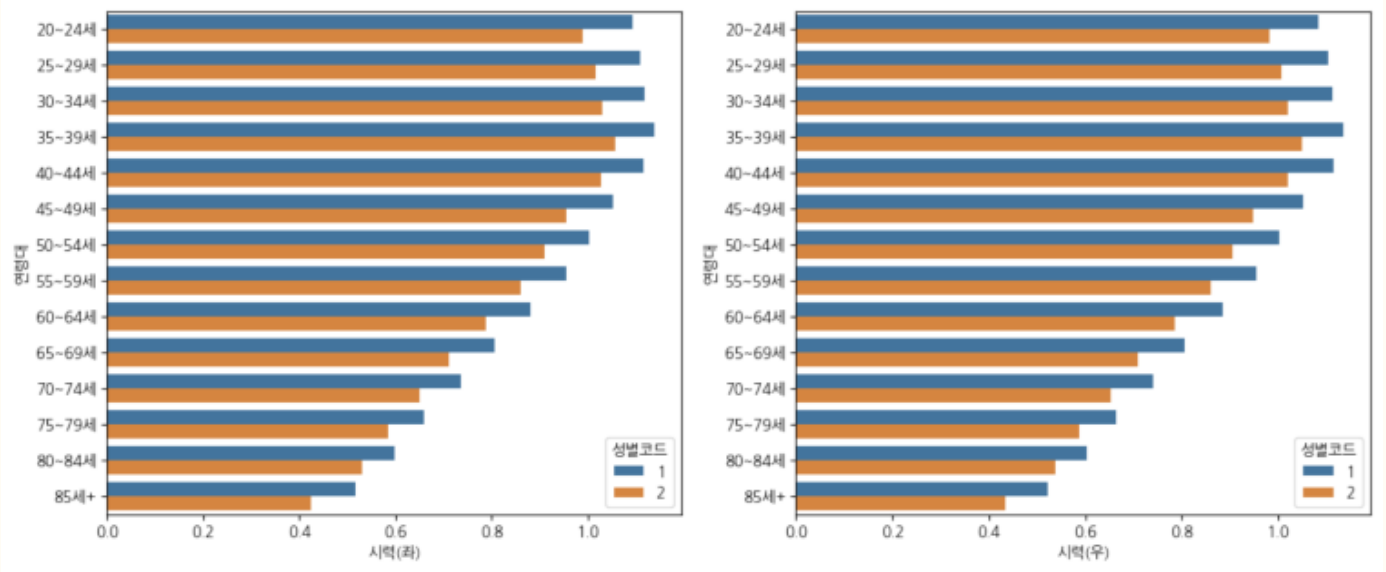


3주차 미션 - 5번

차이 시각화

📌 Q5.

연령대별로 시력은 얼마나 차이가 날까요? 연령대, 성별 좌우 평균 시력을 시각화 해주세요!



코치에게 물어봐

+실시간 QnA

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

솔솔

❓ 아직 학생이어서 RnD 부서가 아닌 다른 부서에서는 통계분석과 머신러닝 기법은 어떻게 사용되는지 궁금합니다.

1.통계분석은 어떨때 사용되는지 궁금합니다

- 제품 성능은 주로 논문을 이용하는 듯하던데 일상 업무에서도 사용하시는지요.
- 사용하신다면 주로 사용하는 통계 기법이 정해져 있는지 궁금합니다.

2. 실무에서는 머신러닝은 어떨 때 활용하시나요?

- 예측 서비스 개발 같은 목표가 있을 때만 사용하는지요.

3. 데이터 분야 직군은 분석보다 ETL 역량을 키워야할까요?

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

Nancy_리더

❓ 일단 python을 잘 하지 못해서 주석을 열심히 달고 있습니다.
(혹시 제가 제 코드를 이해하지 못할까봐요)

공부할때야 제 맘대로 달긴 하지만 이렇게 결과물을 내야할때마다 고민됩니다.
가끔 유튜브에서 영상 보다보면 주석을 다는 것이 좋다 / 불필요한 주석은 필요없다 등등으로 나뉘는거 같더라구요.
주석 설명을 다는 기준이 어떻게 될까요?

코치에게 물어봐

슬랙 #코치에게-물어봐 채널에 남겨주신 질문에 대해 답변해드려요

미구미

❓ 직무에 대한 고민이 많아 짧게나마 적어봅니다...!
수학 전공으로 수치해석, 기계학습을 수강하여 파이썬을 어느정도 이해하고 있습니다.
공공데이터 분석 공모전을 통해 파이썬으로 다룰 수 있는 쉬운 분석과 아이디어 기획도 해본 경험이 있습니다.
그런데 막상 실무적인 활동은 해본 적이 없는 것 같습니다. 데이터 사이언스가 주전공이 아니다보니 이제 막 시작하고 있는 중입니다.
지금부터는 어떤 공부와 경험을 쌓아야 할까요? 그리고 저와 맞는 직무를 찾아가기 위해서는 어떤 활동이 필요할까요?
대학원에서 연구도 해보고 싶고, 직접 실무 환경에서 경험치도 쌓고 싶습니다. 조언 부탁드립니다!

라이브 코칭 4회차

08월 19일 월요일 20시

많은 참여 부탁드려요 🍀

오늘 진행한 3회차 다시보기는 **이번 주 목요일 15시**에 업로드 됩니다:)
리드부스터는 **08월 18일 일요일**까지 활동일지 제출해주세요!