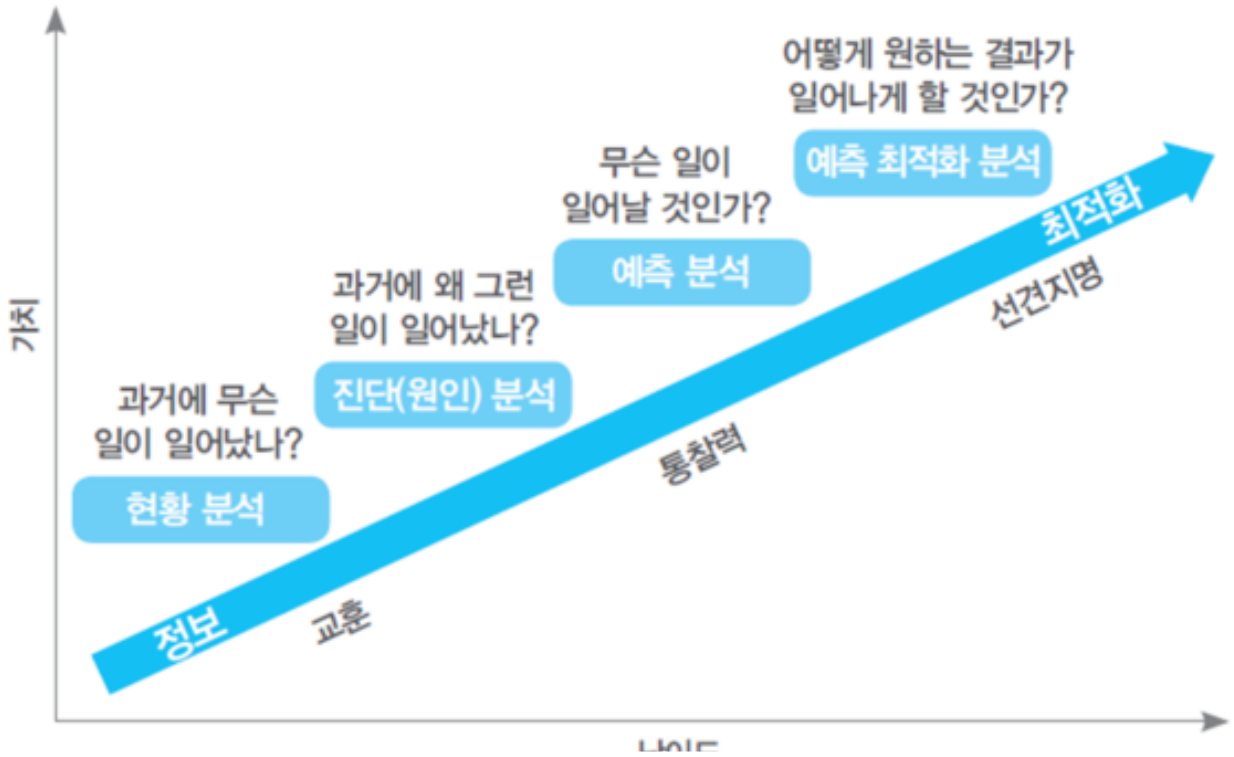
12. 데이터 탐색

탐색적 데이터 분석

- 다양한 각도에서 데이터를 관찰하고 이해하는 과정
  - 각 데이터 특성/분포 확인
  - 데이터 간의 관련성 확인
  - 파생 데이터 생성 및 추가 수집
  - 분석 목표, 가정 수립
  - 분석 및 예측을 위한 데이터 선별



자료: 가트너

데이터 탐색이란 ?

- 데이터 분석을 위해 데이터가 어떤 특징을 가지고 있는 탐색하는 과정
- 개별 데이터(변수) 이해, 데이터간 (변수간) 상관 분석등을 수행
- 또한 시각화를 통해 데이터 이해 및 분석에 대한 통찰력을 획득
- 데이터 전처리 관점
  - 이상 데이터, 결측 데이터를 보정하는 방법 선택
- 데이터 분석 관점
  - 분석에 사용할 주요 변수 (데이터) 선택
  - 분석에 사용할 수 있는 파생 변수 생성, 데이터 변환 등을 수행

데이터 사이의 상관 관계

- 개별 데이터 (예: 키, 몸무게) 를 표현한 것을 변수라고 하였을 때, 두 개의 변수 들이 함께 변화하는 관계를 상관관계 (correlation) 라고 함.
- 산점도, 히스토그램 그래프와 같은 시각화를 통해 두 변수의 상관관계를 파악할 수 있음
- 두 변수 사이의 선형 상관관계 정도를 나타내는 상관계수 (correlation coefficient)
  - 한 변수의 값이 증가할 때 다른 변수의 값도 증가하면, 양의 상관관계가 있다고 하며 반대의 경우는 음의 상관 관계가 있다고 함
  - 양의 상관관계 예시 : 기온이 높으면 아이스크림 판매가 증가
  - 음의 상관관계 예시 : 기온이 높으면 뜨거운 커피 판매가 감소

상관계수 정의

- **상관계수**(correlation coefficient)**는 두 변수 사이의 통계적 관계를 표현하기 위해 특정한 상관 관계의 정도를 수치적으로 나타낸 계수**

- **피어슨 상관 계수**

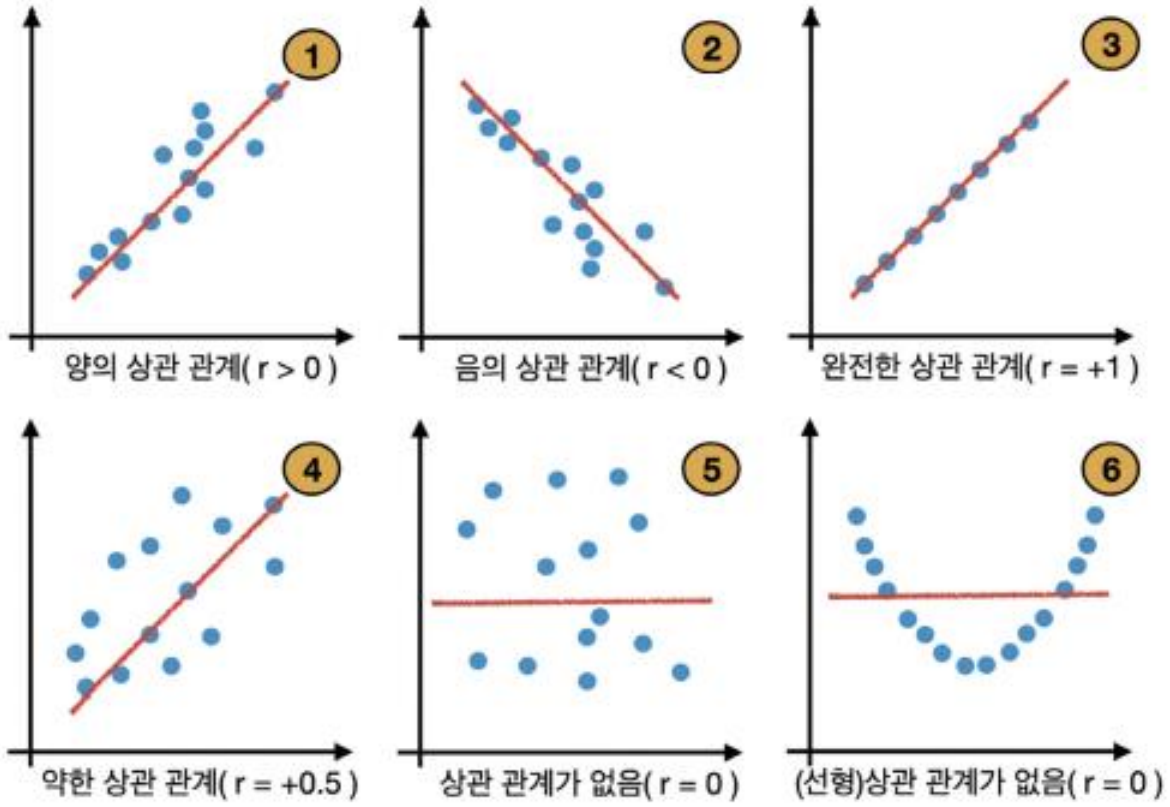
$$\text{피어슨상관계수} = \frac{\text{공분산}}{\text{표준편차} \cdot \text{표준편차}}$$

$$r_{XY} = \frac{\frac{\sum_i^n \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{n-1}}{\sqrt{\frac{\sum_i^n \left(X_i - \overline{X}\right)^2}{n-1}} \sqrt{\frac{\sum_i^n \left(Y_i - \overline{Y}\right)^2}{n-1}}}$$

- df.corr() **메소드에서 상관계수 값을 제공**

산점도 그래프와 상관계수

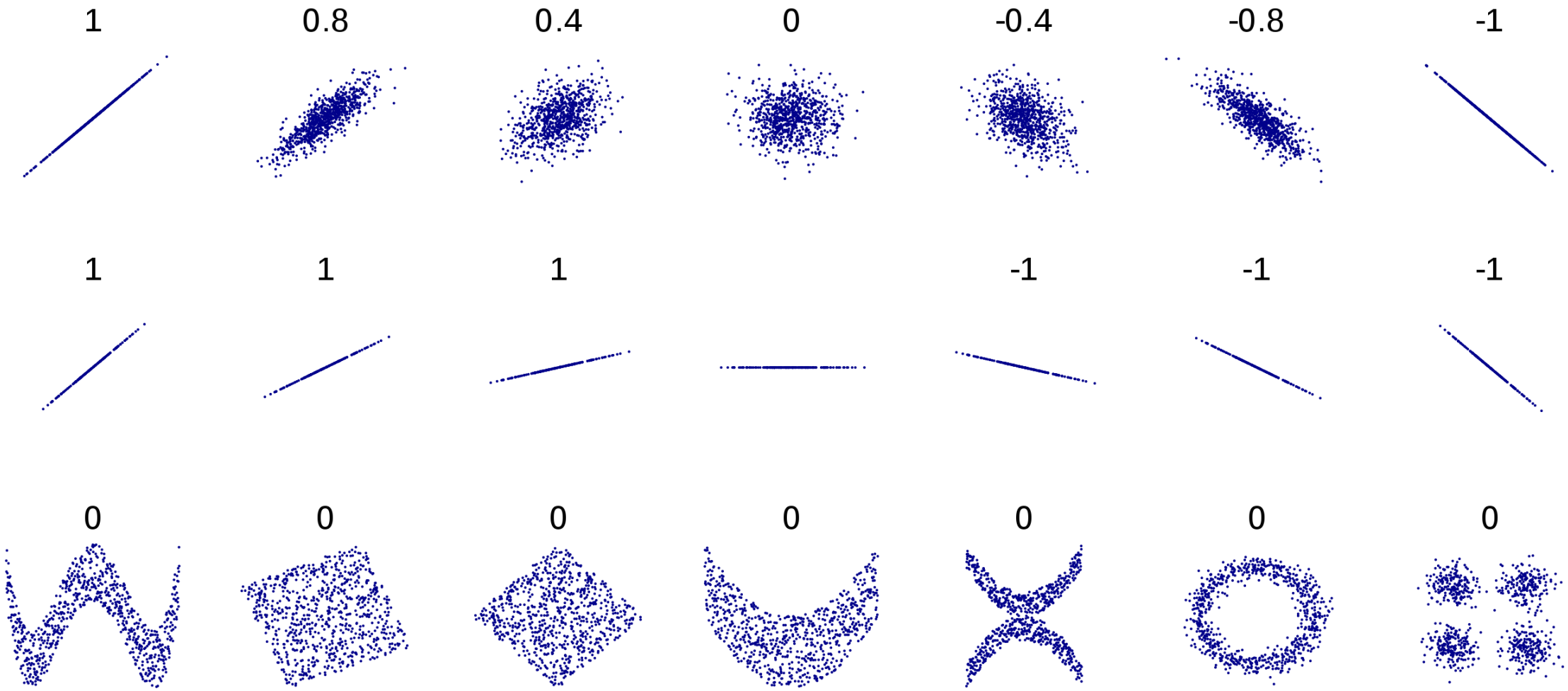
- 상관계수를 r이라고 하고 키와 몸무게와 같은 독립 변수를 각각 x, y 축에 배치한 후 데이터를 산점도 그래프로 시각화하면 다음과 같이 상관 계수를 확인할 수 있음



선형적 관련성을 알려주는 상관계수

- 상관계수는 +1과 -1 사이의 값을 가짐
- 상관계수가 양수인 경우: 한 변수가 증가할 때 다른 한쪽도 증가
- 상관계수가 음수인 경우: 한 변수가 증가할 때 다른 한쪽은 감소
- 상관계수가 +1 인 경우: 두 변수가 완벽한 선형 함수로 표현 가능 (기울기 값은 상관없음)
- 상관계수 값은 분산이 커질수록 작아 짐
- 상관계수가 0 인 경우
  - 데이터의 분포가 랜덤함. (상관관계가 없음)
  - 상관계수는 선형적인 상관도만을 측정할 수 있으며, 비 선형적 관계는 파악하지 못함

상관계수 예시



중앙 그래프는 Y 의 분산이 0 이므로 상관 계수가 정의되지 않는다

데이터 탐색 예시

- 국영수 과목 점수에 대한 관련성을 탐색
  - describe() 결과로는 과목별 특성 및 관련성이 잘 드러나지 않음

```
Korean=[90, 95, 85, 80, 75, 80, 75, 70, 65, 65]
English=[95, 90, 85, 80, 85, 70, 70, 60, 65, 55]
Math=[95, 85, 80, 70, 60, 70, 85, 80, 95, 45]
scores = pd.DataFrame( {'korean':Korean,
                        'english':English, 'math':Math} )
scores.describe()
```

	korean	english	math
count	10.000000	10.000000	10.0000
mean	78.000000	75.500000	76.5000
std	10.055402	13.426756	15.6436
min	65.000000	55.000000	45.0000
25%	71.250000	66.250000	70.0000
50%	77.500000	75.000000	80.0000
75%	83.750000	85.000000	85.0000
max	95.000000	95.000000	95.0000

데이터 탐색 예시

- corr() 메소드를 통해 상관계수 값을 확인
  - 국어와 영어 점수가 상관관계가 높음을 알 수 있음
  - 수학은 국어 & 영어와 상관관계가 낮음을 알 수 있음

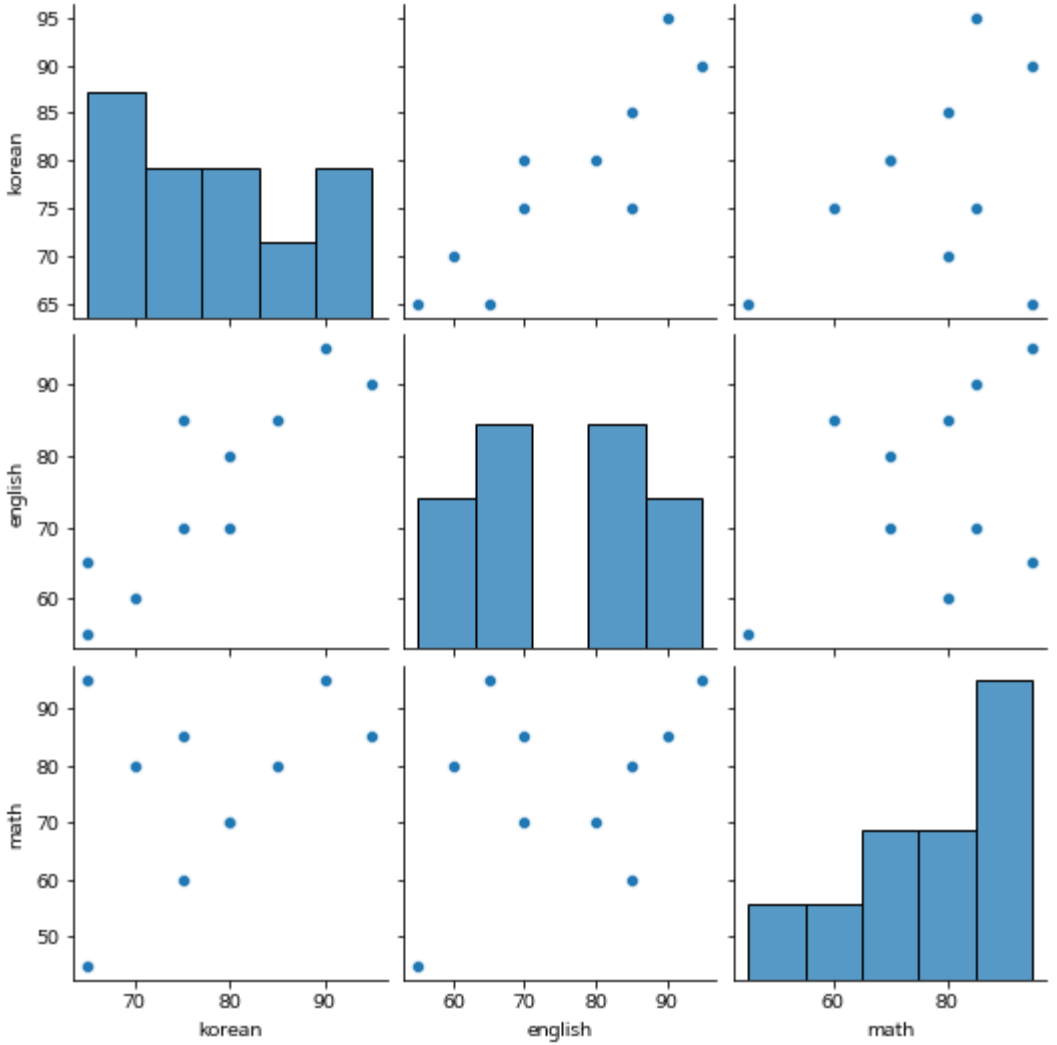
```
print( scores.corr() )
```

	korean	english	math
korean	1.000000	0.872354	0.374367
english	0.872354	1.000000	0.379552
math	0.374367	0.379552	1.000000

데이터 탐색 예시

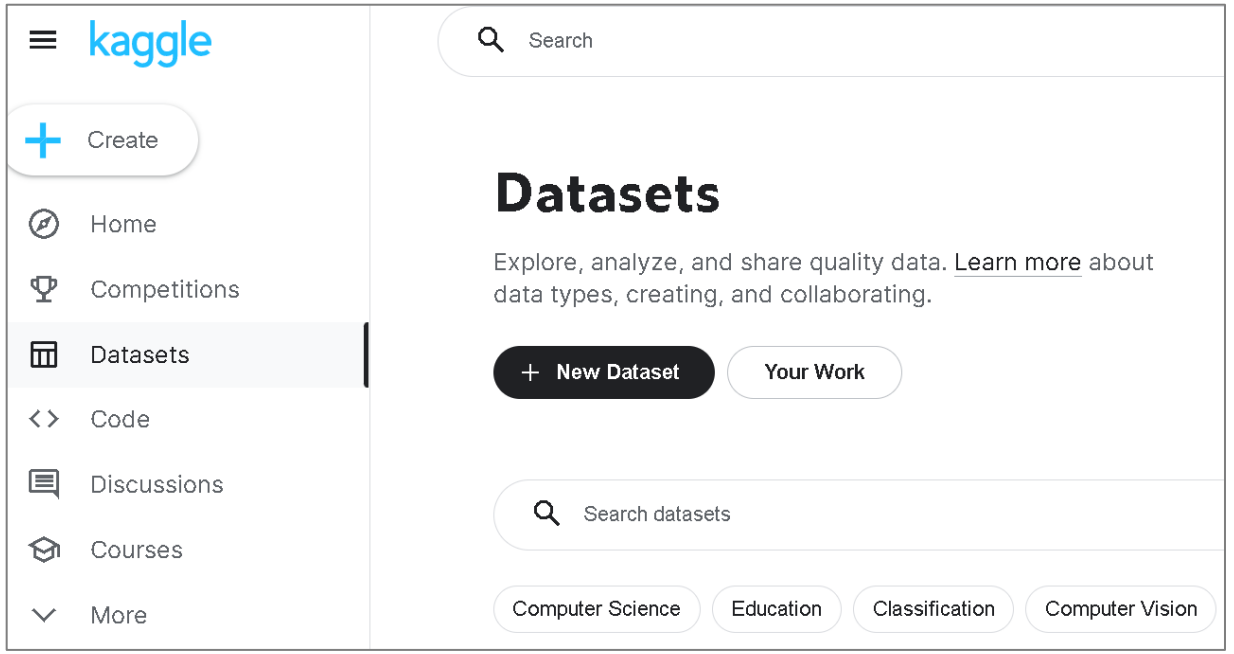
- pairplot() 을 사용하여 데이터 간의 관련성을 시각화
  - 국어와 영어는 상관관계가 높아 보임
  - 국영수 점수가 비례하는 학생도 있으나,
  - 국어 & 영어 점수가 낮아도 수학 점수가 높은 학생이 있음

```
sns.pairplot(data=scores)
plt.show()
```



데이터 분석 도전 : kaggle (캐글)

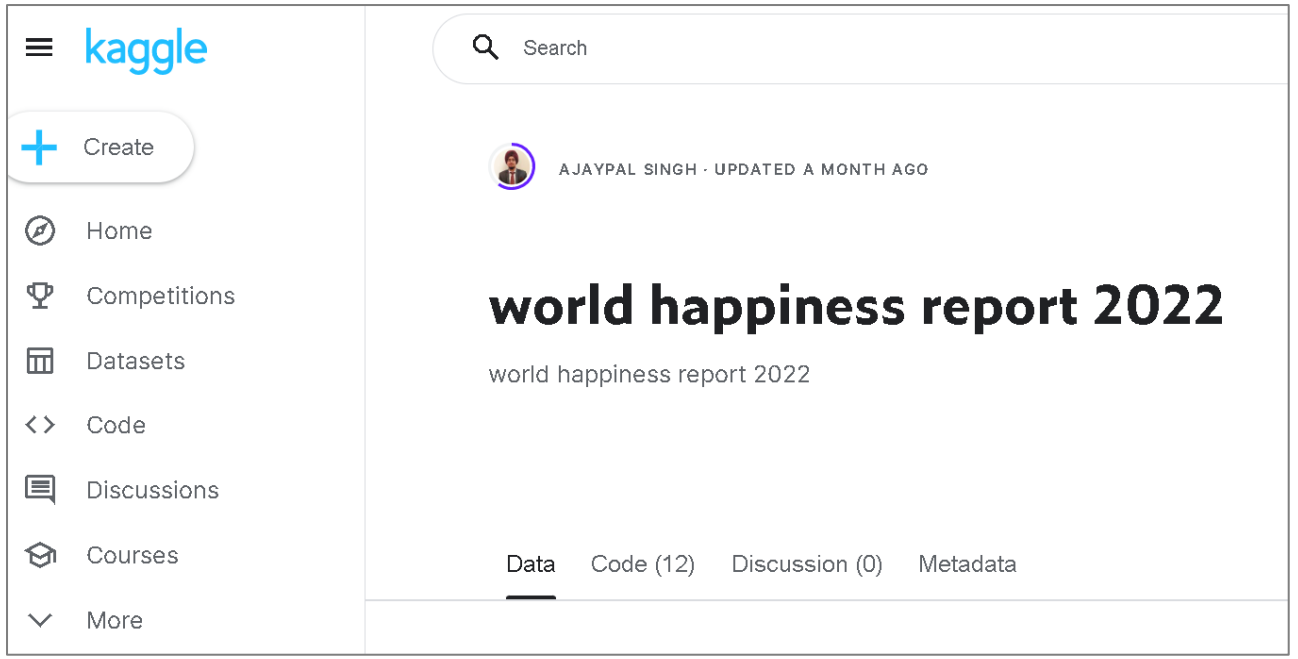
- 캐글은 예측모델 및 분석 대회 플랫폼
- 기업 및 단체에서 데이터와 해결과제를 등록하면, 누구나 자유롭게 참가하여 도전, 경쟁 및 공유
  - Competitions 메뉴를 통해 선의의 경쟁을 하며 이후 코드 공유 문화가 발달
  - Datasets 은 다양한 데이터셋을 축적하고 공유하며 데이터 분석 훈련이 큰 도움



https://www.kaggle.com/

kaggle : World Happiness Report 2022

- 유엔 산하 자문 기구 "지속가능 발전해법 네트워크" 가 매년 발표하는 보고서
  - 세계 각 나라 거주민들의 행복을 정량화 하여 행복지수로 표현
  - 정부, 기업 및 시민 사회에서 활용할 수 있도록 제공



World Happiness Report (WHR)

- 다음과 같은 지표로 행복 지수를 산출함
  - GDP per capita : 1인당 국내 총생산 GDP (소득 수준)
  - Social Support : 어려움에 처했을 때, 도움을 요청할 가족 혹은 친구가 있는지
  - Healthy Life expectancy : 건강한 기대 수명 (WHO 데이터 기준)
  - Freedom to make life choices : 삶의 결정에 대한 충분한 자유가 있는지
  - Generosity : 지난달 기부를 한 적이 있는지
  - Perceptions of corruption : 정부와 기업에 대한 믿음 수준
  - Dystopia + residual : 미래에 대한 불안함 수준

WHR 2022 에서 무엇을 탐색 할 것인가?

- 기본적인 전처리는 되어 있는가? (중복, 결측, 이상 데이터)
- 행복지수 상위/하위 10개 국가는 어디인가?
- 각 지표들의 분포는 어떠한 가?
- 행복지수와 관련성이 높은/낮은 지표는 무엇인가?
- 행복지수 기준 상중하 국가로 나누면 그들 간의 특성은 어떻게 다른 가?
- 그 이외에 흥미로운 정보는 어떤 것이 있는가?

WHR 2022 데이터 전처리

```
from google.colab import drive
drive.mount('/content/drive')
whr = pd.read_csv('/content/drive/MyDrive/etc/whr_2022.csv')
data = whr.copy()
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 146 entries, 0 to 145
Data columns (total 12 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   RANK                                  146 non-null    int64
 1   Country                               146 non-null    object
 2   Happiness score                       146 non-null    float64
 3   Whisker-high                          146 non-null    float64
 4   Whisker-low                           146 non-null    float64
 5   Dystopia (1.83) + residual            146 non-null    float64
 6   Explained by: GDP per capita          146 non-null    float64
 7   Explained by: Social support          146 non-null    float64
 8   Explained by: Healthy life expectancy 146 non-null    float64
 9   Explained by: Freedom to make life choices 146 non-null float64
 10  Explained by: Generosity              146 non-null    float64
 11  Explained by: Perceptions of corruption 146 non-null  float64
dtypes: float64(10), int64(1), object(1)
memory usage: 13.8+ KB
```

#불필요한 항목 제거

```
data2 = data.drop(
    ['Whisker-high', 'Whisker-low' ], axis=1)
```

WHR 2022 데이터 전처리

```
# 컬럼 이름을 간단하게 변경하기
data2.columns = ['rank', 'country', 'happy_score', 'residual', 'gdp',
                 'social_support', 'health', 'freedom', 'generosity', 'trust']

print( data2 )

print( data2.duplicated().sum() ) # 중복데이터 체크

print( data2.isnull().sum() )     # 결측 데이터 체크
```

```
     rank      country  happy_score  residual    gdp  social_support  health  freedom  generosity  trust
0       1      Finland        7.821     2.518  1.892           1.258   0.775    0.736       0.109  0.534
1       2      Denmark        7.636     2.226  1.953           1.243   0.777    0.719       0.188  0.532
2       3      Iceland        7.557     2.320  1.936           1.320   0.803    0.718       0.270  0.191
3       4  Switzerland        7.512     2.153  2.026           1.226   0.822    0.677       0.147  0.461
4       5  Netherlands        7.415     2.137  1.945           1.206   0.787    0.651       0.271  0.419
..    ...          ...          ...       ...    ...             ...     ...      ...         ...    ...
141   142    Botswana*        3.471     0.187  1.503           0.815   0.280    0.571       0.012  0.102
142   143      Rwanda*        3.268     0.536  0.785           0.133   0.462    0.621       0.187  0.544
143   144     Zimbabwe        2.995     0.548  0.947           0.690   0.270    0.329       0.106  0.105
144   145      Lebanon        2.955     0.216  1.392           0.498   0.631    0.103       0.082  0.034
145   146  Afghanistan        2.404     1.263  0.758           0.000   0.289    0.000       0.089  0.005

[146 rows x 10 columns]
```

WHR 2022 데이터 탐색

```
# 행복지수 값으로 정렬
data3 = data2.sort_values(
   'happy_score', ascending=False)

# 행복지수 상위 10 개국
print( data3.head(10) )
print()
# 행복지수 하위 10 개국
print( data3.tail(10) )
```

rank	country	happy_score	residual	gdp	social_support	health	freedom	generosity	trust
1	Finland	7.821	2.518	1.892	1.258	0.775	0.736	0.109	0.534
2	Denmark	7.636	2.226	1.953	1.243	0.777	0.719	0.188	0.532
3	Iceland	7.557	2.320	1.936	1.320	0.803	0.718	0.270	0.191
4	Switzerland	7.512	2.153	2.026	1.226	0.822	0.677	0.147	0.461
5	Netherlands	7.415	2.137	1.945	1.206	0.787	0.651	0.271	0.419
6	Luxembourg*	7.404	2.042	2.209	1.155	0.790	0.700	0.120	0.388
7	Sweden	7.384	2.003	1.920	1.204	0.803	0.724	0.218	0.512
8	Norway	7.365	1.925	1.997	1.239	0.786	0.728	0.217	0.474
9	Israel	7.364	2.634	1.826	1.221	0.818	0.568	0.155	0.143
10	New Zealand	7.200	1.954	1.852	1.235	0.752	0.680	0.245	0.483

rank	country	happy_score	residual	gdp	social_support	health	freedom	generosity	trust
137	Zambia	3.760	1.135	0.930	0.577	0.306	0.525	0.203	0.083
138	Malawi	3.750	1.661	0.648	0.279	0.388	0.477	0.140	0.157
139	Tanzania	3.702	0.735	0.848	0.597	0.425	0.578	0.248	0.270
140	Sierra Leone	3.574	1.556	0.686	0.416	0.273	0.387	0.202	0.055
141	Lesotho*	3.512	1.312	0.839	0.848	0.000	0.419	0.076	0.018
142	Botswana*	3.471	0.187	1.503	0.815	0.280	0.571	0.012	0.102
143	Rwanda*	3.268	0.536	0.785	0.133	0.462	0.621	0.187	0.544
144	Zimbabwe	2.995	0.548	0.947	0.690	0.270	0.329	0.106	0.105
145	Lebanon	2.955	0.216	1.392	0.498	0.631	0.103	0.082	0.034
146	Afghanistan	2.404	1.263	0.758	0.000	0.289	0.000	0.089	0.005

WHR 2022 데이터 탐색

```
# 'rank' 및 'country' 열은 제거
data3.drop( ['rank', 'country'],
 axis=1, inplace=True )

print( data3.head(10).mean() )
print()
print( data3.tail(10).mean() )
print()

# 행복지수 상위 & 하위 국가 비교
print( data3.head(10).mean() /
data3.tail(10).mean() )
```

 # 상위 10개 평균

```
happy_score          7.4658
residual             2.1912
gdp                  1.9556
social_support       1.2307
health               0.7913
freedom              0.6901
generosity           0.1940
trust                0.4137
dtype: float64
```

 # 하위 10개 평균

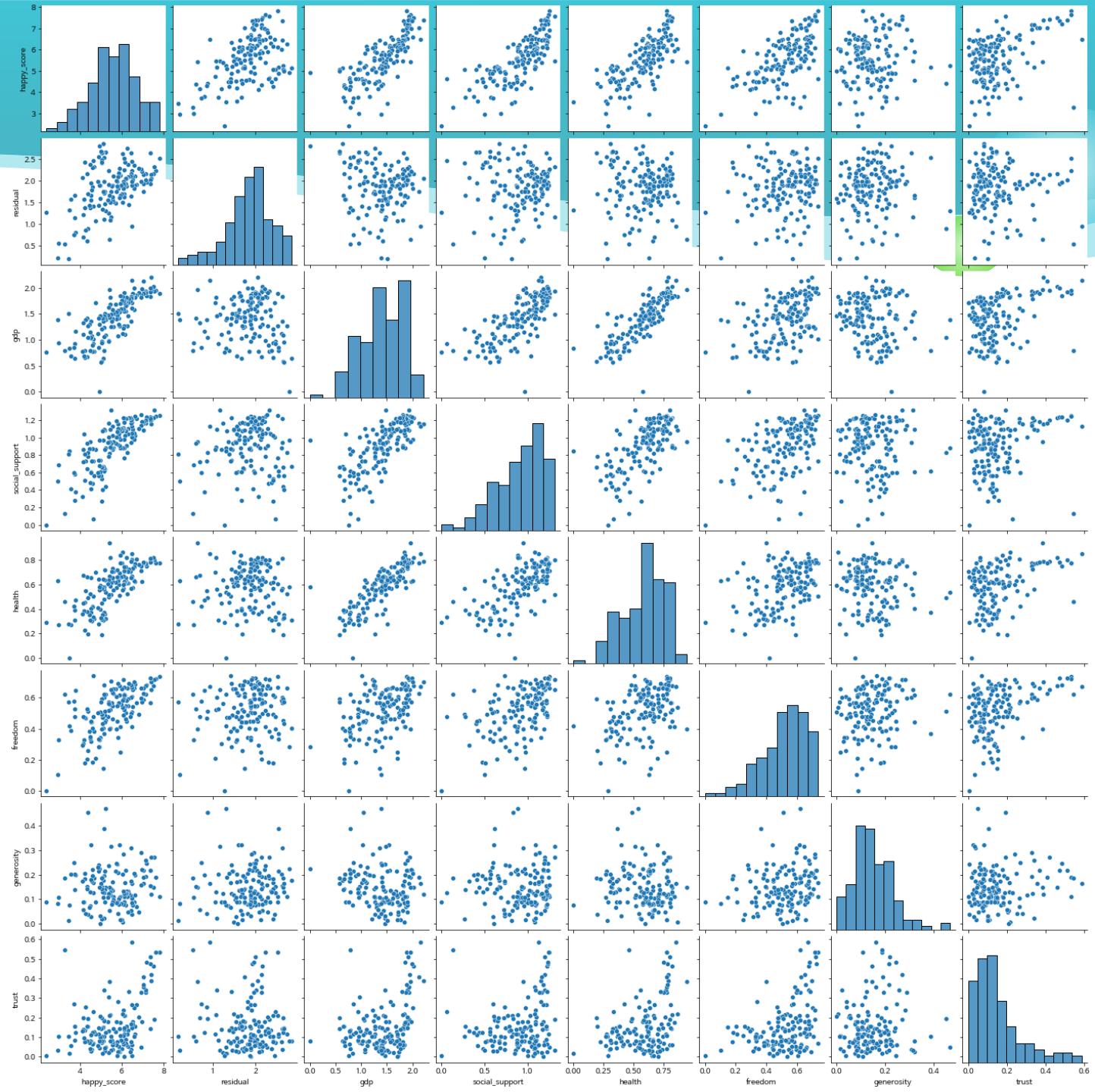
```
happy_score          3.3391
residual             0.9149
gdp                  0.9336
social_support       0.4853
health               0.3324
freedom              0.4010
generosity           0.1345
trust                0.1373
dtype: float64
```

 # 상위 평균 / 하위 평균

```
happy_score          2.235872
residual             2.395016
gdp                  2.094687
social_support       2.535957
health               2.380566
freedom              1.720948
generosity           1.442379
trust                3.013110
dtype: float64
```

- pairplot( ) : 모든 수치형 데이터 간의 산점도 그래프

```
sns.pairplot(data3)
plt.show()
```



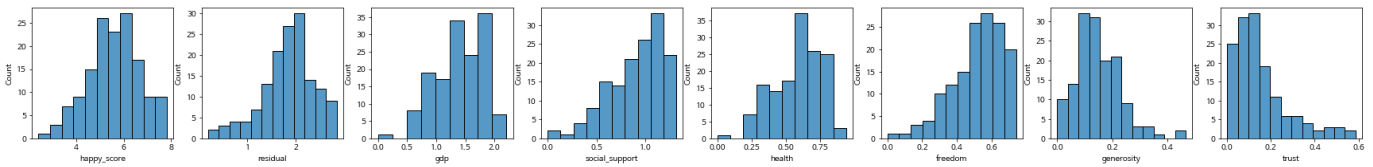
WHR 2022 데이터 탐색

- 반복문을 사용하여 각 열에 대해 boxplot / histplot 등을 시각화 할 수 있음
- subplot 을 사용하여 원하는 레이아웃으로 시각화 할 수 있음

```
for c in data3.columns:
  plt.figure ( figsize = (4, 3))
  sns.boxplot(x=c, data=data3)
  plt.show()
```

```
for c in data3.columns:
  plt.figure ( figsize = (4, 3))
  sns.histplot(x=c, data=data3)
  plt.show()
```

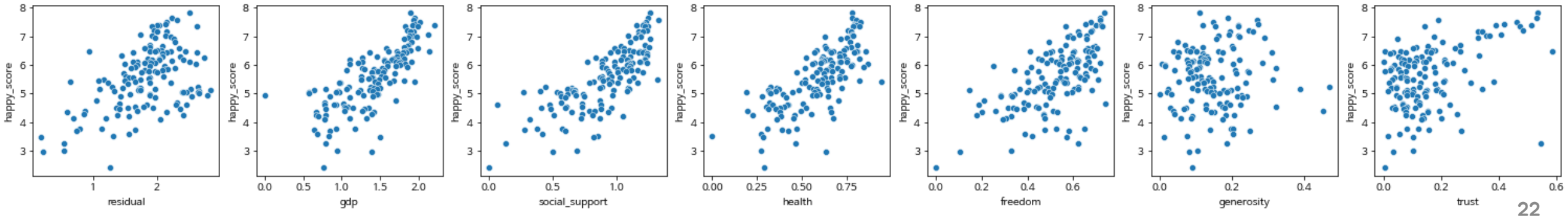
```
plt.figure ( figsize = (30, 3))
i=1
for c in data3.columns:
  #plt.subplot(1, len(data3.columns), i )
  sns.histplot(x=c, data=data3)
  i = i+1
  sns.boxplot(x=c, data=data3)
plt.show()
```



WHR 2022 데이터 탐색 : 행복지수 중점으로

- 행복지수 값과 다른 데이터간의 산점도 그래프로 관련성을 시각화

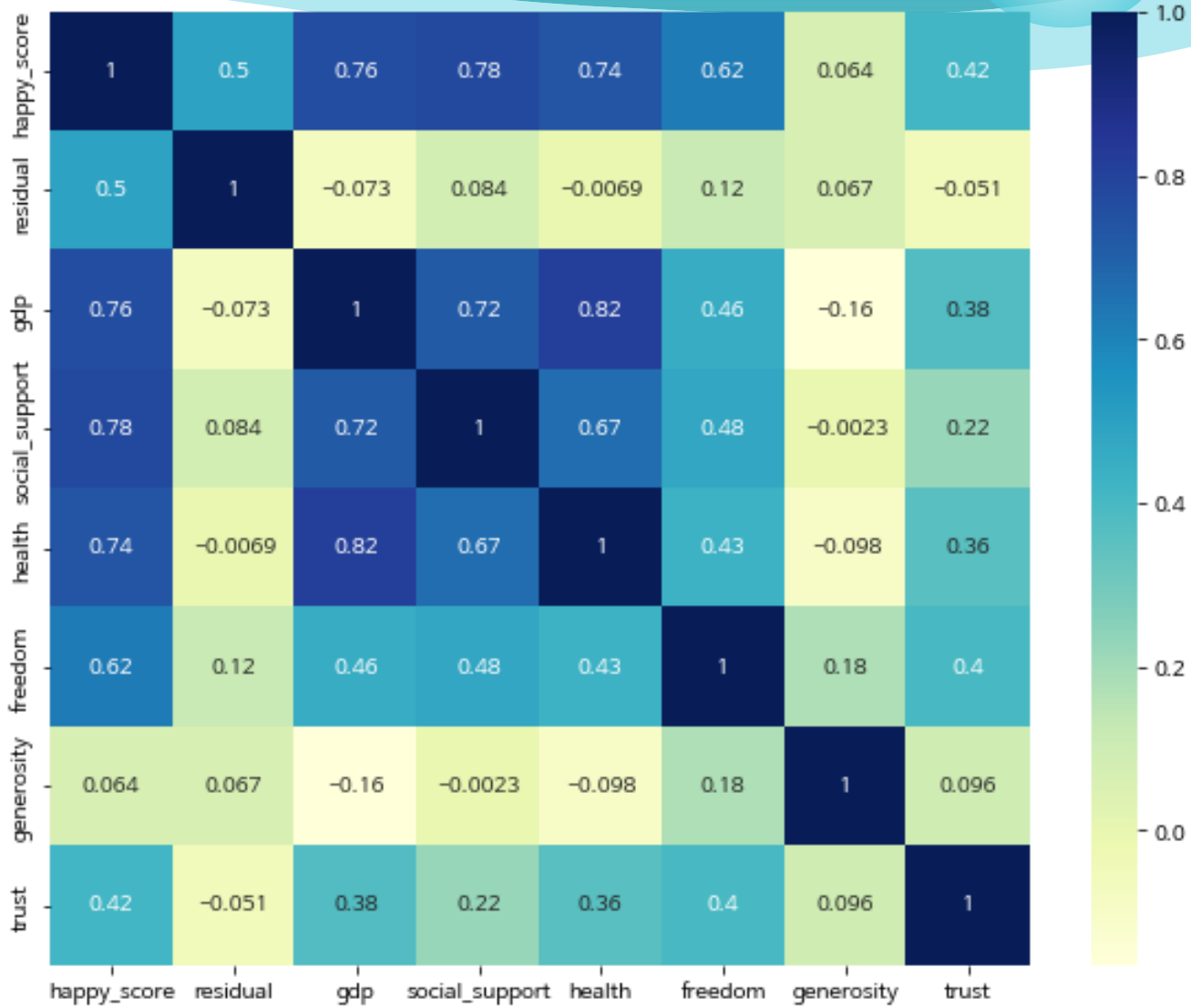
```
plt.figure ( figsize = (30, 3))
i=1
for c in data3.columns:
  if c != 'happy_score':
    plt.subplot(1, len(data3.columns)-1, i )
    i = i+1
    sns.scatterplot(x=c, y='happy_score', data=data3)
plt.show()
```



WHR 2022 데이터 탐색 : 상관계수

- df.corr() 메소드 사용

```
corr_res = data3.corr()
print(corr_res)
plt.figure( figsize=(10, 8))
sns.heatmap(corr_res, annot=True, cmap='YlGnBu')
plt.show()
```

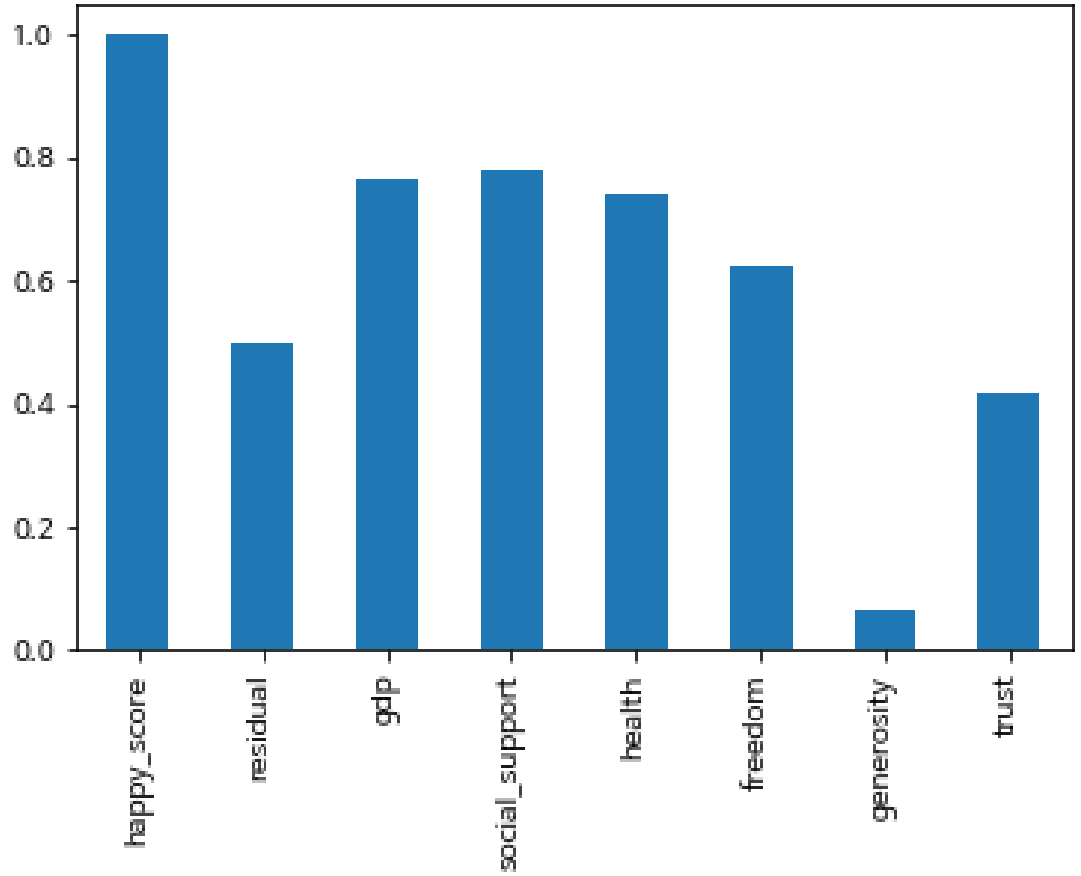


WHR 2022 데이터 탐색 : 상관계수

- corr() 결과 중 행복지수 것 만 가져와서 시각화

```
print( corr_res.loc['happy_score'] )
corr_res.loc['happy_score'].plot.bar()
plt.show()
```

```
happy_score          1.000000
residual             0.498990
gdp                  0.763677
social_support       0.777889
health               0.740260
freedom              0.624822
generosity           0.063785
trust                0.416216
Name: happy_score, dtype: float64
```



WHR 2022 데이터 탐색 : 파생 변수 추가

- 행복지수 기준으로 상중하 그룹 정보 열을 생성

```
# 행복지수 순위 1/3, 2/3 지점 값으로
# 상, 중, 하 국가로 그루핑 함
def encoding_group_rank(x):
  if x >= h_border:
    return 'H'
  elif x >= m_border:
    return 'M'
  else:
    return 'L'

data3['group_rank'] = data3['happy_score'].apply( encoding_group_rank )
```

```
# 행 개수 (나라 개수)
nums = data3.shape[0]
# 행복지수로 정렬되어 있으므로 1/3 지점, 2/3 지점의
# 행복지수 값을 가져옴
h_border = data3.iloc[ int(nums/3) ]['happy_score']
m_border = data3.iloc[ int(nums/3) * 2 ]['happy_score']
print( h_border, m_border )
```

WHR 2022 데이터 탐색 : 파생 변수 추가

- 행복지수 상중하 그룹 별 평균을 구하여 비교하자

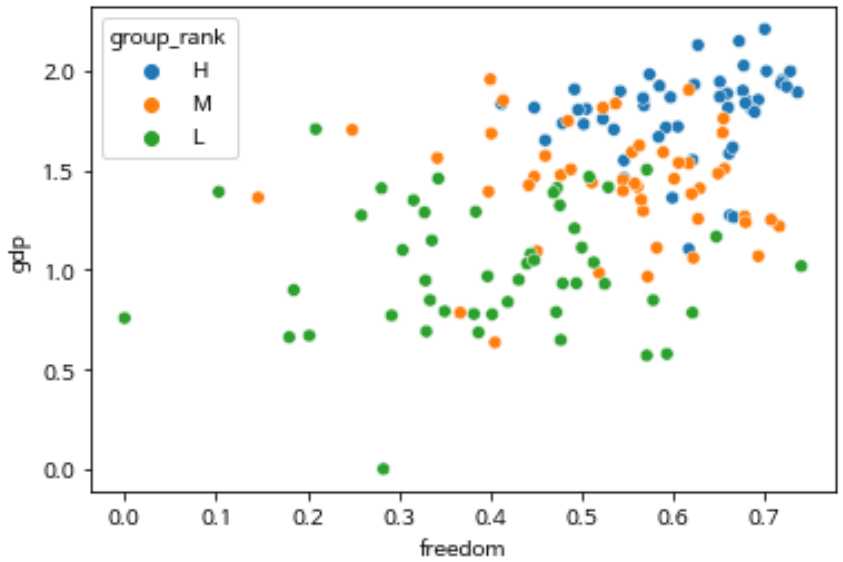
```
data3.groupby('group_rank').mean().sort_values('happy_score', ascending=False)
```

group_rank	happy_score	residual	gdp	social_support	health	freedom	generosity	trust
H	6.711204	2.080980	1.789449	1.131143	0.721816	0.612204	0.149531	0.226163
M	5.597313	1.789542	1.428333	0.960167	0.619542	0.535229	0.145750	0.118813
L	4.353102	1.624041	1.013918	0.627388	0.417837	0.404612	0.146816	0.118633

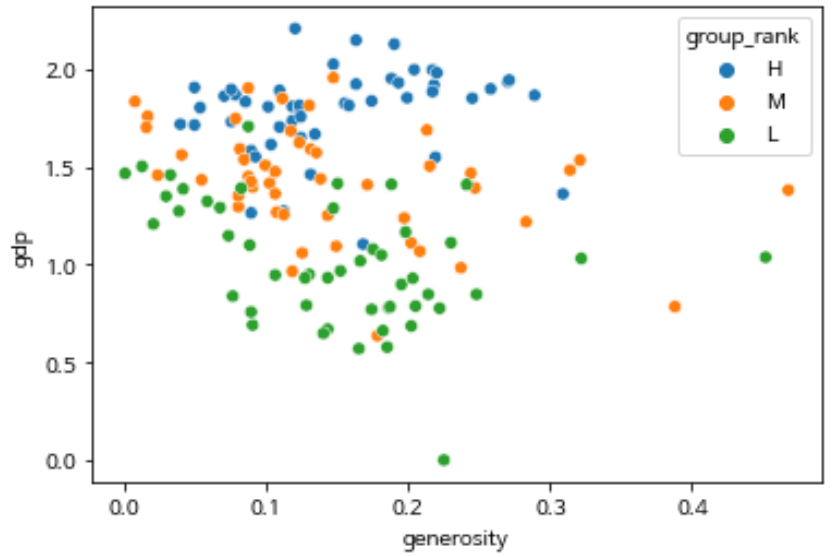
WHR 2022 데이터 탐색 : 그 외의 흥미로운 정보

- (행복지수 상중하 정보와 함께) GDP 와 다른 지표 간의 관련성은 ?

```
sns.scatterplot( x='freedom', y='gdp',
hue='group_rank', data=data3 )
plt.show()
```



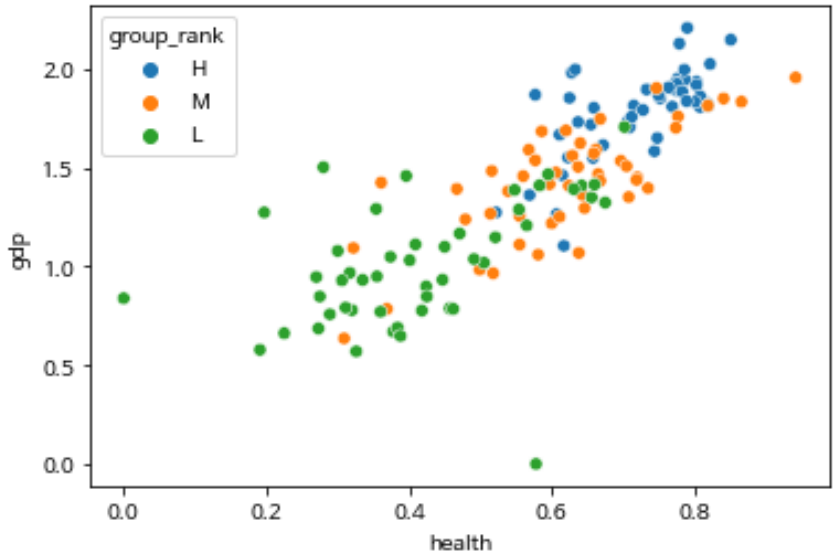

```
sns.scatterplot( x='generosity', y='gdp',
 hue='group_rank', data=data3 )
plt.show()
```



WHR 2022 데이터 탐색 : 그 외의 흥미로운 정보

- (행복지수 상중하 정보와 함께) GDP 와 다른 지표 간의 관련성은 ?

```
sns.scatterplot( x='health', y='gdp',
 hue='group_rank', data=data3 )
plt.show()
```




```
sns.scatterplot( x='trust', y='gdp',
 hue='group_rank', data=data3 )
plt.show()
```