IT 파이썬. 딥러닝

데이터 분석

강사 이수현

목차

데이터 분석

- 전처리 과정
- 데이터 불러오기
- 데이터 분석 및 시각화

문제 정의하기

문제 상황 이해하기

나라별 행복 지수와 관련 있는 것은 무엇일까요?

탐색할 정보 알아보기

데이터 분석 과정에서 탐색할 정보

[탐색 정보1] 우리나라의 행복 지수는 몇 위일까요? 그리고 나라별 행복 지수 순위는 어떻게 될까요?

행복 지수가 높은 나라 순서 또는 낮은 나라 순서로 정렬하여 순서를 확인 하기

문제 정의하기

탐색할 정보 알아보기

데이터 분석 과정에서 탐색할 정보를 확인

[탐색 정보2] 전 세계 나라별 행복 지수를 한눈에 보기 쉽게 표현하는 방법을 찾아볼까요?

반응형 그래프(Interactive graph)를 사용하여 시각화(플로틀리(plotly) 사용).

treemap, sunburst, choropleth 기법을 사용해 보면서 반응형 시각화 기법

이를 통해 다양한 시각화 기법이 있다는 것을 알 수 있습니다.

문제 정의하기

탐색할 정보 알아보기

데이터 분석 과정에서 탐색할 정보를 미리 살펴봅시다.

[탐색 정보3] 행복 지수 속성과 관련 깊은 속성은 무엇일까요?

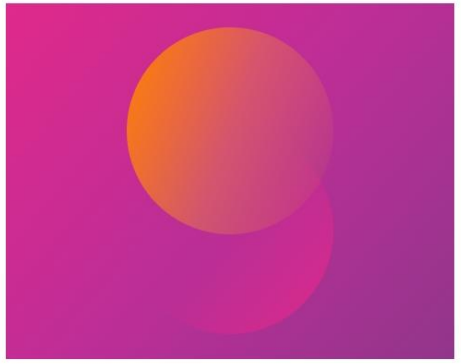
행복 지수 속성과 다른 속성 간의 상관관계를 분석하여 나온 상관계수 값으로 행복 지수 속성과 관련 깊은 속성을 확인

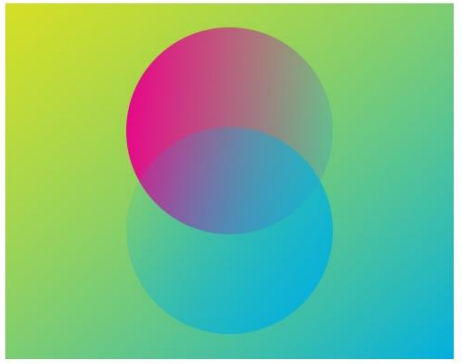
데이터 불러오기

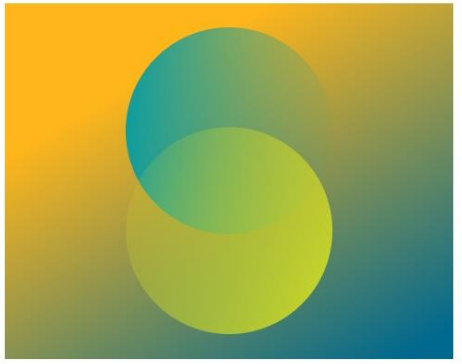
데이터셋 소개

유엔 산하 자문기구인 지속가능발전 해법 네트워크에서 발표하는 세계 행복 보고서를 기반으로 해며, 국가명, 지역, 나라별 행복 지수, 1인당 국내총생산 등 다양한 속성이 있는 데이터이다.

World Happiness Report







데이터 불러오기

행복 지수 데이터 내려받기

캐글 검색창에 'world happness report' 검색하여 아래 데이터를 다운로드

World Happiness Report

Happiness scored according to economic production, social support, etc.

Data Card | Code (1046) | Discussion (16)

About Dataset

Context

The World Happiness Report is a landmark survey of the state of global happiness. The first report was published in 2012, the second in 2013, the third in 2015, and the fourth in the 2016 Update. The World Happiness 2017, which ranks 155 countries by their happiness levels, was released at the United Nations at an event celebrating International Day of Happiness on March 20th. The report continues to gain

Usability ⓘ

8.53

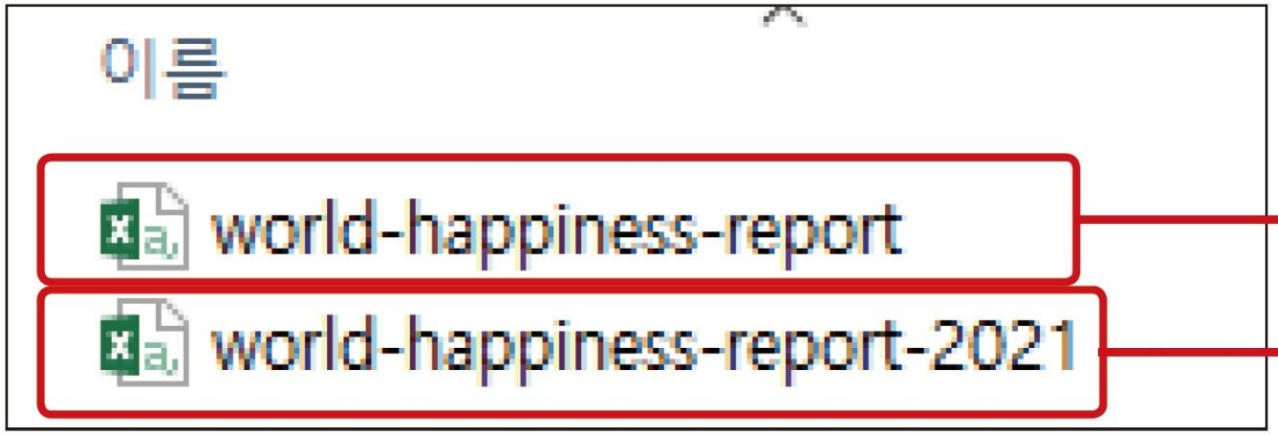
License

CC0: Public Domain

Expected update frequency

행복 지수 데이터 내려받기

압축을 풀면 2개의 데이터 셋을 얻을 수 있으며, 이중 'world-happiness-report-2021.csv.'를 주로 사용한다.



world happiness report:
2006~2020년 연도별 행복 지수

world happiness report 2021:
2021년 행복 지수

행복 지수 데이터 내려받기

'world-happiness-report-2021.csv.' 파일 확인

1	Country name	Regional indicator	Ladder score	Standard err	upperwhisl	lowerwhis	Logged GDP	Social sup
2	Finland	Western Europe	7.842	0.032	7.904	7.78	10.775	0.954
3	Denmark	Western Europe	7.62	0.035	7.687	7.552	10.933	0.954
4	Switzerland	Western Europe	7.571	0.036	7.643	7.5	11.117	0.942
5	Iceland	Western Europe	7.554	0.059	7.67	7.438	10.878	0.983
6	Netherlands	Western Europe	7.464	0.027	7.518	7.41	10.932	0.942
7	Norway	Western Europe	7.392	0.035	7.462	7.323	11.053	0.954
8	Sweden	Western Europe	7.363	0.036	7.433	7.293	10.867	0.934
9	Luxembourg	Western Europe	7.324	0.037	7.396	7.252	11.647	0.908
10	New Zealand	North America and ANZ	7.277	0.04	7.355	7.198	10.643	0.948

데이터 불러오기

데이터셋불러오기

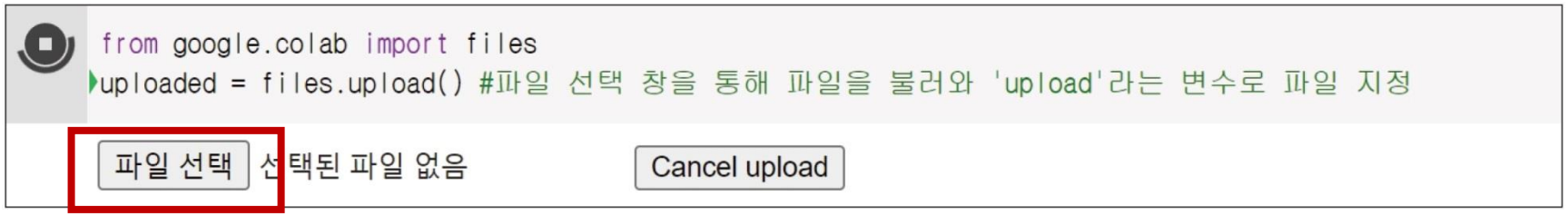
데이터를 다루기 위한 판다스(pandas), 처리한 데이터를 시각화하기 위한 맷플롯립(matplotlib), 시본(seaborn), 플로틀리(plotly) 라이브러리 추가

```
1  import pandas as pd
2  import matplotlib.pyplot as plt
3  import seaborn as sns
4  import plotly.express as px)
```

데이터 불러오기

파일 업로드하기

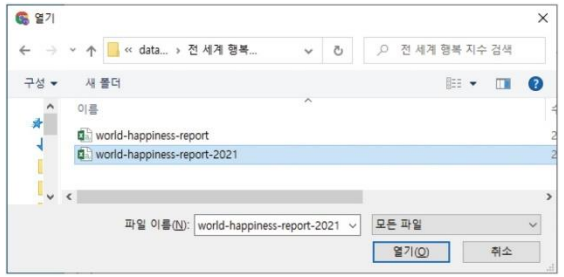
google. colab 라이브러리를 사용해 파일을 업로드



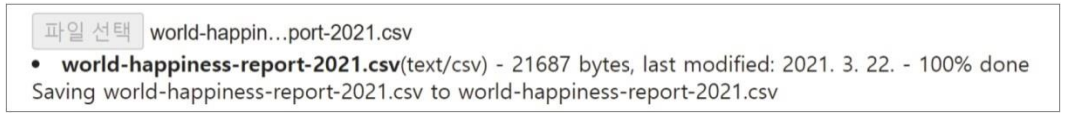
파일 업로드하기

코랩에 파일을 업로드하기

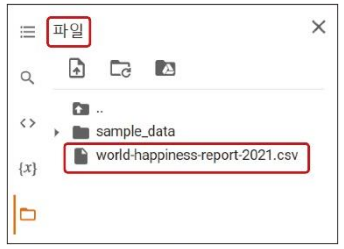
① 파일 선택 버튼을 클릭하여 파일을 업로드한다.



② 업로드 완료 후 안내 메시지 출력을 확인한다.



③ 업로드된 파일이 화면 왼쪽 파일에 저장된 것을 확인한다.



데이터 불러오기

판다스로 파일 읽어 들이기

판다스 라이브러리의 read_csv( ) 함수를 사용해 코랩 노트북으로 파일을 읽어 와서 사용해 파일을 업로드

```
데이터프레임 객체 = 판다스 객체.read_csv(경로 변수)
# 판다스 라이브러리의 read_csv( ) 함수를 사용하여
파일을 코랩 노트북으로 읽어 들임.
```

판다스로 파일 읽어 들이기

읽어 들인 파일을 데이터프레임 형태로 출력한다.

```
data_path = 'world-happiness-report-2021.csv' # 변수에 파일 경로 저장
happiness_data = pd.read_csv(data_path)
happiness_data
```

	Country name	Regional indicator	Ladder score	Standard error of ladder score	upperwhisker	lowerwhisker	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption	Ladder score in Dystopia	Explained by: Log GDP per capita	Explained by: Social suppor
0	Finland	Western Europe	7.842	0.032	7.904	7.780	10.775	0.954	72.000	0.949	-0.098	0.186	2.43	1.446	1.106
1	Denmark	Western Europe	7.620	0.035	7.687	7.552	10.933	0.954	72.700	0.946	0.030	0.179	2.43	1.502	1.108
2	Switzerland	Western Europe	7.571	0.036	7.643	7.500	11.117	0.942	74.400	0.919	0.025	0.292	2.43	1.566	1.079
3	Iceland	Western Europe	7.554	0.059	7.670	7.438	10.878	0.983	73.000	0.955	0.160	0.673	2.43	1.482	1.172
4	Netherlands	Western Europe	7.464	0.027	7.518	7.410	10.932	0.942	72.400	0.913	0.175	0.338	2.43	1.501	1.079
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

데이터 탐색 및 시각화하기

python™

데이터 살펴보기

문제 해결을 위해 데이터를 살펴보고 해결 가능한 형태로 전처리

데이터 기초 정보 확인하기

판다스 라이브러리의 info() 메소드를 사용하여 데이터의 기초 정보를 확인

```
데이터프레임 객체.info( )
# info( ) 메소드를 통해 데이터 개수, 속성 개수, 속성명, 결측치,
속성의 데이터 유형 등 확인
```

데이터 탐색 및 시각화하기

데이터 기초 정보 확인하기

happiness_data.info( )

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 149 entries, 0 to 148
Data columns (total 20 columns):
 #   Column                                Non-Null Count  Dtype
---  ------                                --------------  -----
 0   Country name                          149 non-null    object
 1   Regional indicator                    149 non-null    object
 2   Ladder score                          149 non-null    float64
 3   Standard error of ladder score        149 non-null    float64
 4   upperwhisker                          149 non-null    float64
 5   lowerwhisker                          149 non-null    float64
 6   Logged GDP per capita                 149 non-null    float64
 7   Social support                        149 non-null    float64
 8   Healthy life expectancy               149 non-null    float64
 9   Freedom to make life choices          149 non-null    float64
 10  Generosity                            149 non-null    float64
 11  Perceptions of corruption             149 non-null    float64
 12  Ladder score in Dystopia              149 non-null    float64
 13  Explained by: Log GDP per capita      149 non-null    float64
 14  Explained by: Social support          149 non-null    float64
 15  Explained by: Healthy life expectancy 149 non-null    float64
 16  Explained by: Freedom to make life choices 149 non-null float64
 17  Explained by: Generosity              149 non-null    float64
 18  Explained by: Perceptions of corruption 149 non-null  float64
 19  Dystopia + residual                   149 non-null    float64
dtypes: float64(18), object(2)
memory usage: 23.4+ KB
```

데이터 탐색 및 시각화하기

데이터 기초 정보 확인하기

이 데이터 셋은 총 149개의 데이터로 구성되어 있으며, 속성은 20개, 결측치는 없음(non-null)을 확인

속성별 데이터 유형은 실수형(float64) 18개, 문자열(object) 2개로 구성

데이터 탐색 및 시각화하기

python

주요 속성 추출하기

전체 데이터 중 일부 속성을 추출하기 위해 판다스의 iloc[] 메소드를 사용

```
데이터프레임 객체.iloc[추출할 행 인덱스, 추출할 열 인덱스]
데이터프레임 객체.iloc[0:5] # 상위 5개 행과 모든 열
데이터프레임 객체.iloc[:, 0:2] # 모든 행과 첫 2개 열
데이터프레임 객체.iloc[[0, 3, 6, 24], [0, 5, 6]] # [0, 3, 6, 24] 행과 [0, 5, 6] 열만
데이터프레임 객체.iloc[0:5, 5:8] # 상위 5개 행과 [5, 6, 7] 열만
```

주요 속성 추출하기

- 각 속성명 설명: 주요 속성

인덱스	주요 속성명	설명
0	Country name	국가명
1	Regional indicator	지역
2	Ladder score	행복 지수
6	Logged GDP per capita	1인당 국내총생산
7	Social support	사회적 지원
8	Healthy life expectancy	건강 수명
9	Freedom to make life choices	삶에 대한 선택의 자유
10	Generosity	관용
11	Perceptions of corruption	부정부패 인식 지수

데이터 탐색 및 시각화하기

주요 속성 추출하기

- **각 속성명 설명: 부가 속성**

인덱스	주요 속성명
3	Standard error of ladder score
4	upperwhisker
5	lowerwhisker
12	Ladder score in Dystopia
13	Explained by: Log GDP per capita
14	Explained by: Social support
15	Explained by: Healthy life expectancy
16	Explained by: Freedom to make life choices
17	Explained by: Generosity
18	Explained by: Perceptions of corruption
19	Dystopia + residual

데이터 탐색 및 시각화하기

주요 속성 추출하기

iloc[] 메소드를 사용하여 주요 속성 9개를 추출한 후, 상위 5개 행까지 출력한다.

```
happiness = happiness_data.iloc[:, [0, 1, 2, 6, 7, 8, 9, 10, 11]] # 주요 속성 인덱스
happiness.head(5) # head( ) 함수의 기본값은 50이므로 5는 생략 가능
```

	Country name	Regional indicator	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
0	Finland	Western Europe	7.842	10.775	0.954	72.0	0.949	-0.098	0.186
1	Denmark	Western Europe	7.620	10.933	0.954	72.7	0.946	0.030	0.179
2	Switzerland	Western Europe	7.571	11.117	0.942	74.4	0.919	0.025	0.292
3	Iceland	Western Europe	7.554	10.878	0.983	73.0	0.955	0.160	0.673
4	Netherlands	Western Europe	7.464	10.932	0.942	72.4	0.913	0.175	0.338

데이터 탐색 및 시각화하기

데이터 통계치 살펴보기

주요 속성의 통계량을 파악하기 위해 describe() 메소드를 사용

happiness.describe( )

	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
count	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000	149.000000
mean	5.532839	9.432208	0.814745	64.992799	0.791597	-0.015134	0.727450
std	1.073924	1.158601	0.114889	6.762043	0.113332	0.150657	0.179226
min	2.523000	6.635000	0.463000	48.478000	0.382000	-0.288000	0.082000
25%	4.852000	8.541000	0.750000	59.802000	0.718000	-0.126000	0.667000
50%	5.534000	9.569000	0.832000	66.603000	0.804000	-0.036000	0.781000
75%	6.255000	10.421000	0.905000	69.600000	0.877000	0.079000	0.845000
max	7.842000	11.647000	0.983000	76.953000	0.970000	0.542000	0.939000

실행 결과를 통해 전체 나라의 수(count)는 149개이며, 행복 지수(Ladder score)의 최댓값은 7.842000, 평균은 5.532839, 최솟값은 2.523000인 것을 알 수 있습니다.

데이터 탐색 및 시각화하기

python

탐색정보미리보기

우리나라의 행복 지수 순위와 나라별 행복 지수 순위를 시각화하여 한눈에 확인

속성 기준으로 데이터 정렬하기

sort_values() 메소드를 사용하여 행복 지수(Ladder score) 속성을 기준으로 데이터를 나라별로 정렬

데이터프레임 객체.sort_values(by = '정렬 기준이 되는 속성명', ascending = True/False)

데이터프레임 객체.loc[추출할 행, 추출할 열] # 열을 생략하면 전체 열을 가져옴.

데이터 탐색 및 시각화하기

속성 기준으로 데이터 정렬하기

행복 지수(Ladder score) 속성을 기준으로 내림차순 정렬 후, 국가명(Country name)이 South Korea인 행을 찾아 출력

```
happiness = happiness.sort_values(by = 'Ladder score', ascending = False)
2 happiness.loc[happiness['Country name'] == 'South Korea']
```

	Country name	Regional indicator	Ladder score	Logged GDP per capita	Social support	Healthy life expectancy	Freedom to make life choices	Generosity	Perceptions of corruption
61	South Korea	East Asia	5.845	10.651	0.799	73.9	0.672	-0.083	0.727

우리나라의 행복 지수(Ladder score)는 5.845이고, 행복 지수 순위는 149개국 중 62위(61번째 인덱스)인 것을 확인

시본 사용하여 시각화하기

시본(Seaborn) 라이브러리란?

- 맷플롯립을 편리하게 사용하기 위해 만든 라이브러리
- 맷플롯립과 완벽하게 호환
- 맷플롯립에 비해 코드가 짧고 결과가 보기 좋음.
- 맷플롯립보다 세부 수정하기가 까다로움.

데이터 탐색 및 시각화하기

시본 사용하여 시각화하기

- **가로형 막대그래프 출력**

가로형 막대그래프를 출력하는 barplot()을 사용

```
sns.barplot(x = 'x축 이름', y = 'y축 이름', data = 데이터프레임 객체, palette = '색상')
# barplot( )의 괄호 안에 x, y축 이름 및 그래프 색상 정보 설정
```

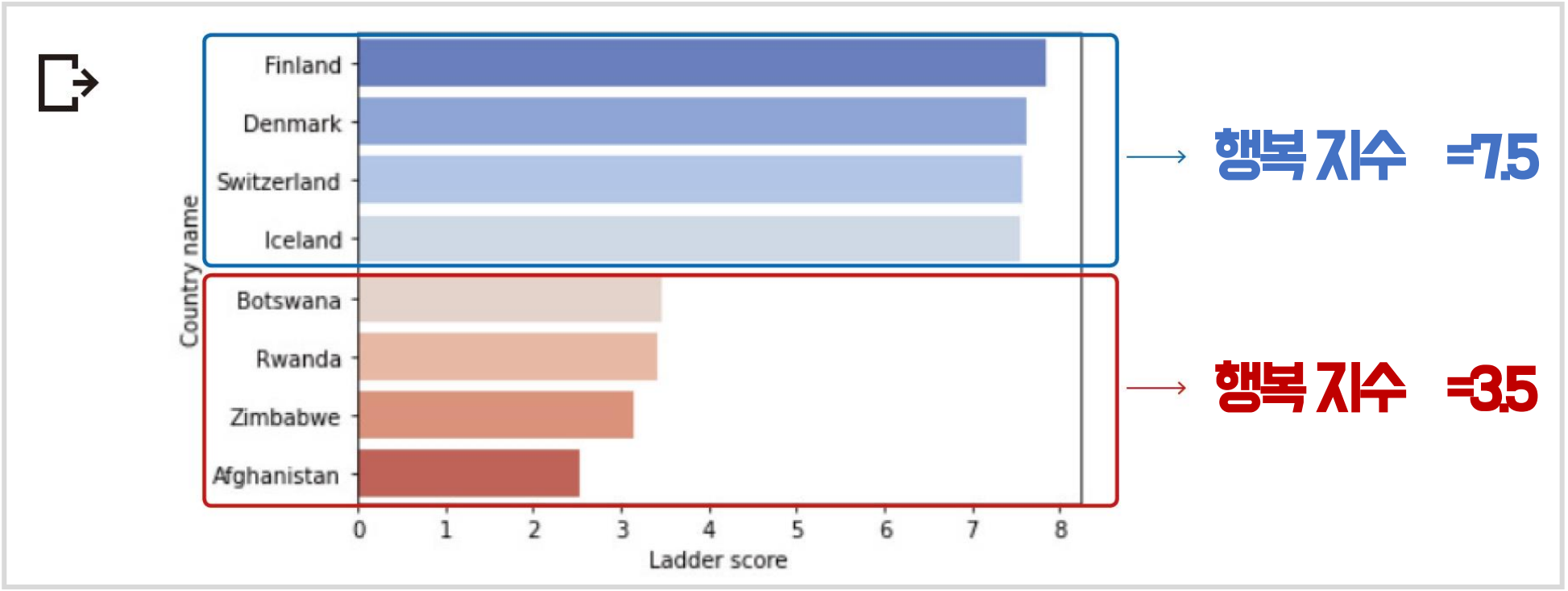
데이터 탐색 및 시각화하기

시본 사용하여 시각화하기

행복 지수 속성값이 7.5 이상인 나라와 3.5 이하인 나라를 가로형 막대그래프로 출력

```
happinessFilter = (happiness.loc[:, 'Ladder score']>= 7.5) |₩
                  happiness.loc[:, 'Ladder score']<= 3.5)
sns.barplot(x = 'Ladder score', y = 'Country name',
            data = happiness[happinessFilter], palette = 'coolwarm')
```

데이터 탐색 및 시각화하기



실행 결과를 통해 행복 지수가 7.5 이상인 나라는 그래프에서 상위 4개 나라이고, 행복 지수가 3.5 이하인 나라는 하위 4개 나라인 것을 확인할 수 있습니다.

데이터 탐색 및 시각화하기

시본 더 알아보기

- **지역(Regional indicator)별 나라의 개수 시각화**

  시본의 countplot() 메소드를 사용하여 지역별 나라의 개수를 시각화해 보자.

데이터 탐색 및 시각화하기

시본 더 알아보기

- **지역(Regional indicator)별 행복 지수의 분포 시각화**

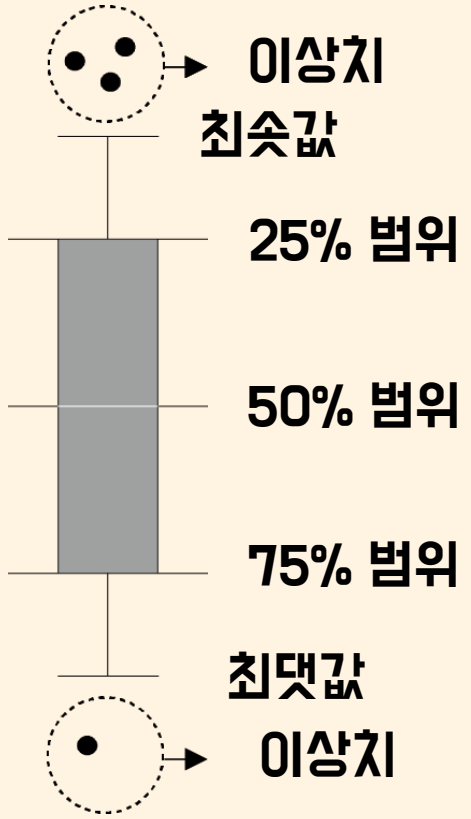
  시본의 boxplot() 메소드를 사용하여 행복 지수의 분포를 시각화해 보자.

```
sns.boxplot(x = '속성명', y = '속성명', data = 데이터프레임 객체, orient = 'h/v')
# x축과 y축은 속성명, data는 데이터프레임 객체, orient는 방향(h는 가로, v는 세로)
```

데이터 탐색 및 시각화하기

시본 더 알아보기

boxplot 시각화로 행복 지수의 25%, 50%, 75%의 분포 범위를 한눈에 확인할 수 있습니다.

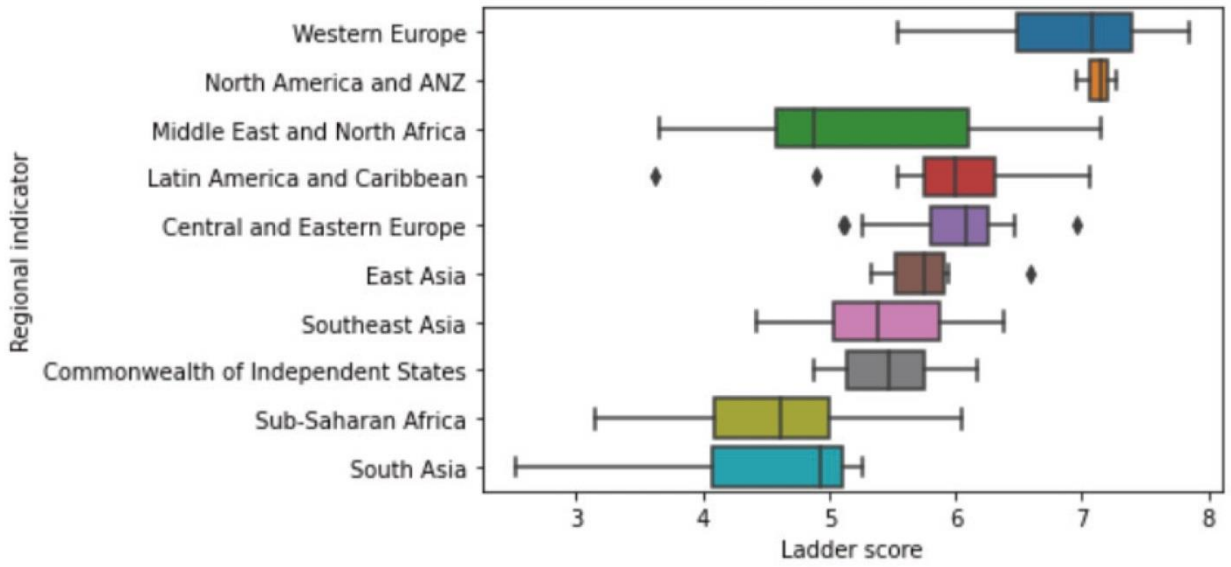


데이터 탐색 및 시각화하기

사본 더 알아보기

각 지역별 행복 지수의 분포도를 출력

```
sns.boxplot(x = 'Ladder score', y = 'Regional indicator', data = happiness, orient = 'h')
```



실행 결과를 통해 주로 Western Europe, North America and ANZ 지역에 행복 지수가 높은 나라가 많이 분포

Sub-Saharan Africa, South Asia 지역에 행복 지수가 낮은 나라가 많이 분포

데이터 탐색 및 시각화하기

python

해 보기

다음 조건에 맞는 boxplot을 출력해 보고, 위 boxplot과 비교해 보자.

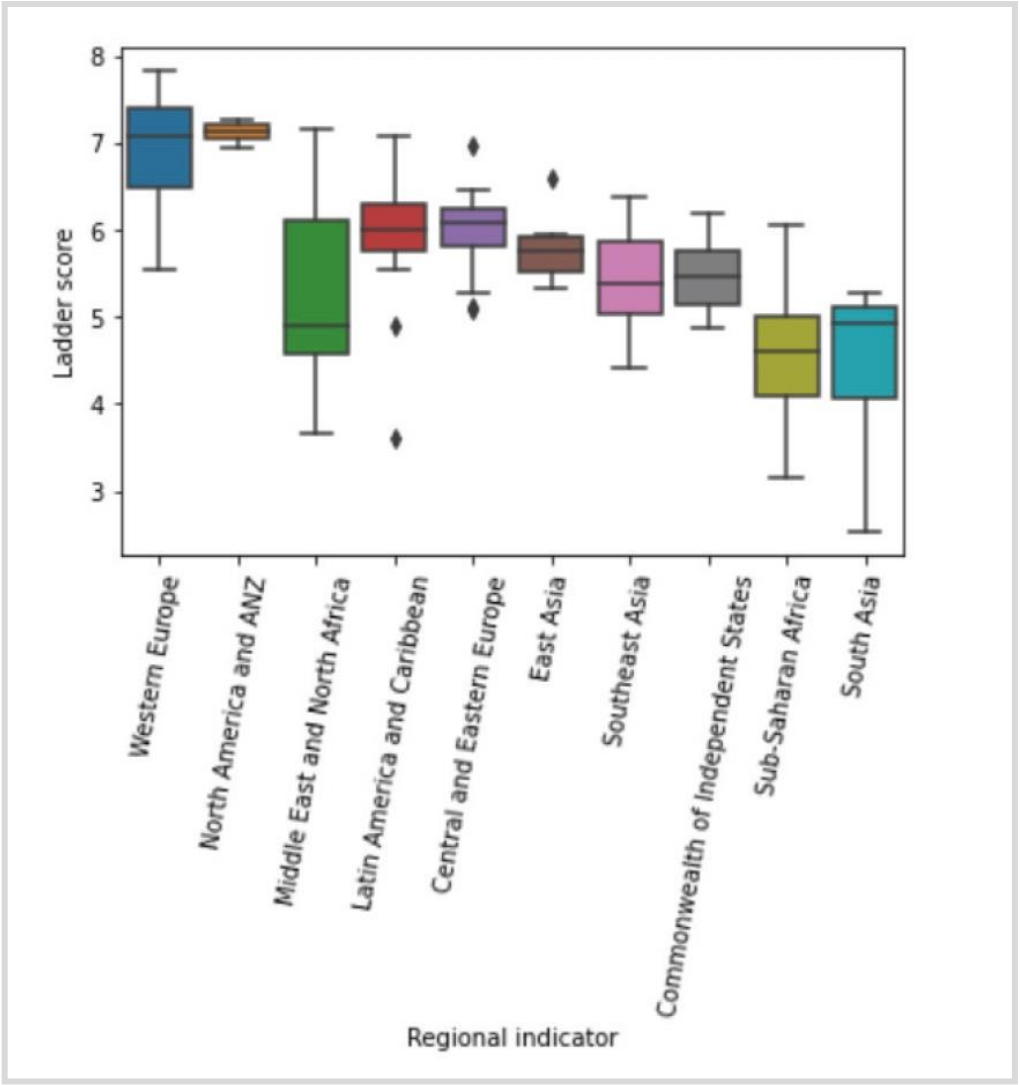
조건

- x축과 y축의 속성값을 바꾸고, boxplot을 세로로 출력한다.
- 레이블을 회전하는 코드를 추가하여 x축의 레이블이 서로 겹치지 않게 한다.

데이터 탐색 및 시각화하기

해 보기

```
sns.boxplot(x = 'Regional indicator',
            y = 'Ladder score',)
            data = happiness, orient = 'v')
 plt.xticks(rotation = 80)
```



데이터 탐색 및 시각화하기

탐색정보2
알아보기

전 세계 나라별 행복 지수를 한눈에 보기 위해 반응형 그래프(Interactive graph)를 사용해 보자.

반응형 그래프(Interactive graph)란?

- 데이터를 계층에 따라 다양한 색으로 표현 가능
- 마우스 움직임에 따라 반응하여 실시간으로 형태가 변하거나 세부 정보를 보여 줄 수 있음.
- treemap(), sunburst(), choropleth() 등

데이터 탐색 및 시각화하기

python™

플로틀리 사용하여 시각화하기

- **treemap 시각화 기법 사용**
  - 1991년 미국의 컴퓨터 과학자인 벤 슈나이더맨(Ben Shneiderman)이 고안한 시각화 방식
  - 계층(트리 구조)을 이루는 데이터 전체와 일부분 간의 관계 파악
  - 범주 간의 정확한 비교보다 큰 특징을 살펴볼 때 주로 사용

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

px.treemap(data_frame = 데이터프레임 객체, path = [부모 열, 자식 열],
values = 열 속성, color = 열 속성)

path는 [부모, 자식] 순서로 작성, values는 사각형 크기가 나타내는 속성,
color는 색상으로 표현하는 속성

데이터 탐색 및 시각화하기

python

플로틀리 사용하여 시각화하기

나라별 행복 지수를 크게 지역별로 구분하여 반응형 그래프로 시각화해 보자. (부모 속성: '지역', 자식 속성: ' 국가명')

```
fig = px.treemap(data_frame = happiness, path = ['Regional indicator',
                                'Country name'], values = 'Ladder score', color = 'Ladder score')
fig.show( ) # plotly 라이브러리에서 그래프 출력
```

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

실행 결과를 통해 전 세계 나라별 행복 지수를 대략적으로 한눈에 비교할 수 있으며, 세부 속성에 마우스를 가져다 대면 지역이나 국가명, 행복 지수 등의 정보 확인

데이터 탐색 및 시각화하기

python™

해 보기

행복 지수가 높은 Western Europe, North America and ANZ와 행복 지수가 낮은 Sub-Saharan Africa, South Asia에는 각각 어떤 나라가 속해 있는지 확인해 보자.

실행 결과인 treemap에서 지역 속성(부모)를 클릭하면 해당하는 지역의 국가명 속성(자식)이 나타납니다.

데이터 탐색 및 시각화하기

해 보기

행복 지수가 높은 Western Europe, North America and ANZ와 행복 지수가 낮은 Sub-Saharan Africa, South Asia에는 각각 어떤 나라가 속해 있는지 확인해 보자.

데이터 탐색 및 시각화하기

해 보기

North America and ANZ를 클릭하면 New Zealand, Australia, Canada, United States의 4개의 나라를 확인할 수 있다.

해 보기

앞 코드에 px.Constant('world')를 추가한 후 실행 결과를 비교해 보자.

```
fig = px.treemap(data_frame = happiness, path = [px.Constant('world'),
                 'Regional indicator', 'Country name'],
                 values = 'Ladder score', color = 'Ladder score')
fig.show( )
```

데이터 탐색 및 시각화하기

해 보기

위 실행 결과를 통해 가장 상위에 'world'가 생성된 것을 확인하였다.

데이터 탐색 및 시각화하기

python

해 보기

앞 코드에 fig.update_layout()메소드를 사용하여 그래프에 제목을 넣는 코드를 추가해 보자.

```
fig.update_layout(title = '제목', title_x = 정렬 위치, width = 너비, height = 길이)
# 그래프의 제목, 정렬 위치(축 방향으로 0.5이면 가운데, 1이면 끝 부분에 정렬),
너비와 길이는 그래프 출력 영역
```

```
fig.update_layout(title = '나라별 행복 지수', title_x = 0.5, width = 900, height = 900)
```

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

- **sunburst 시각화 기법 사용**
  - treemap과 마찬가지로 계층(트리 구조)을 이루는 데이터를 표현하는 데 적합한 시각화 기법식
  - 가장 안쪽에 있는 원이 계층 구조의 부모(계층 구조의 상위), 바깥쪽 원이 자식(계층 구조의 하위)에 해당

px.sunburst(data_frame = 데이터프레임 객체, path = [부모 열, 자식 열],
values = 열 속성, color = 열 속성)

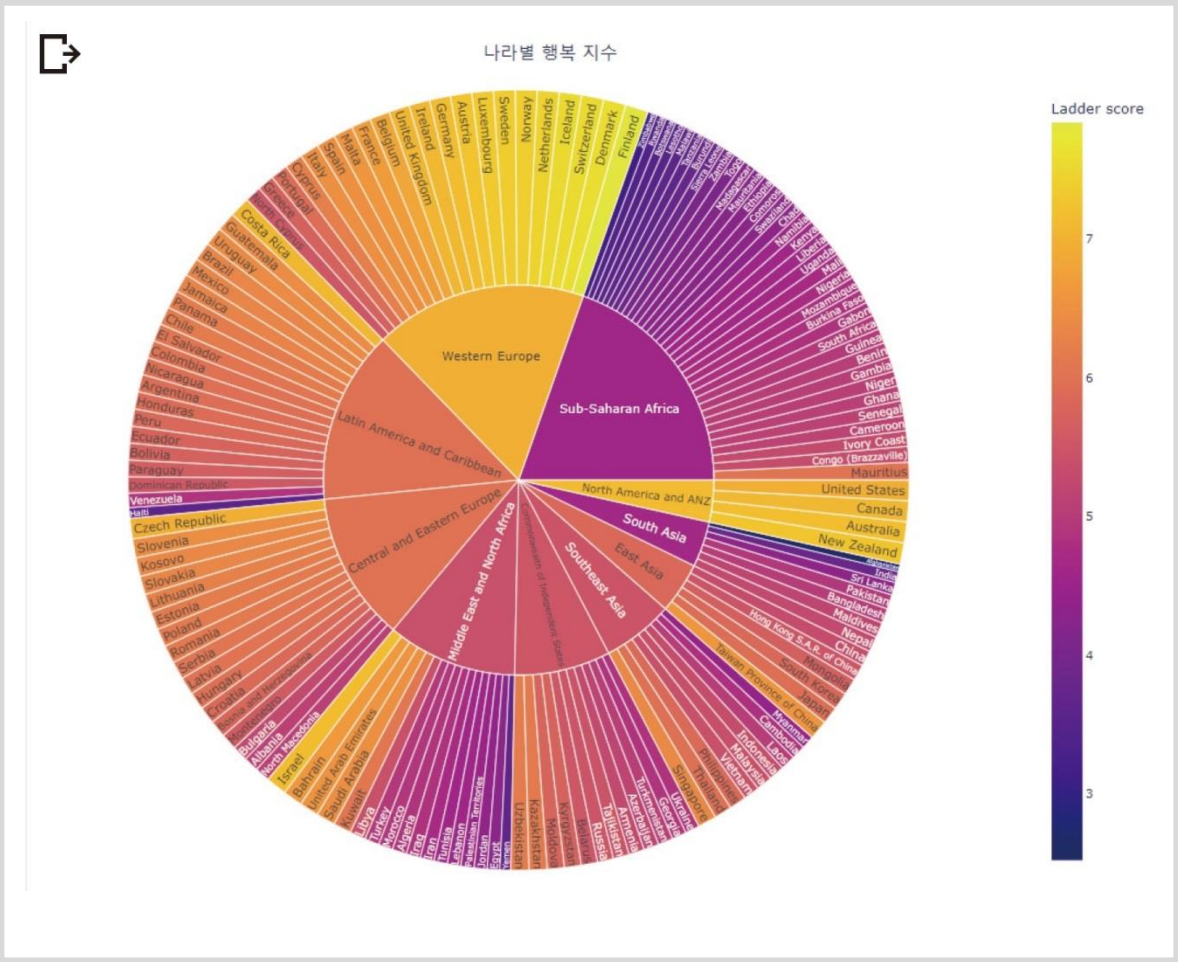
플로틀리 사용하여 시각화하기

다음 코드를 실행한 후 213쪽에서 treemap을 사용했을 때와 sunburst를 사용했을 때 실행 결과를 비교해 보자.

```
1 fig = px.sunburst(data_frame = happiness,
2                   path = ['Regional indicator', 'Country name'],)
3                   values = 'Ladder score', color = 'Ladder score')
4 fig.update_layout(title = '나라별 행복 지수', title_x = 0.5,
5                   width = 1200, height = 900),
6 fig.show( )
```

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

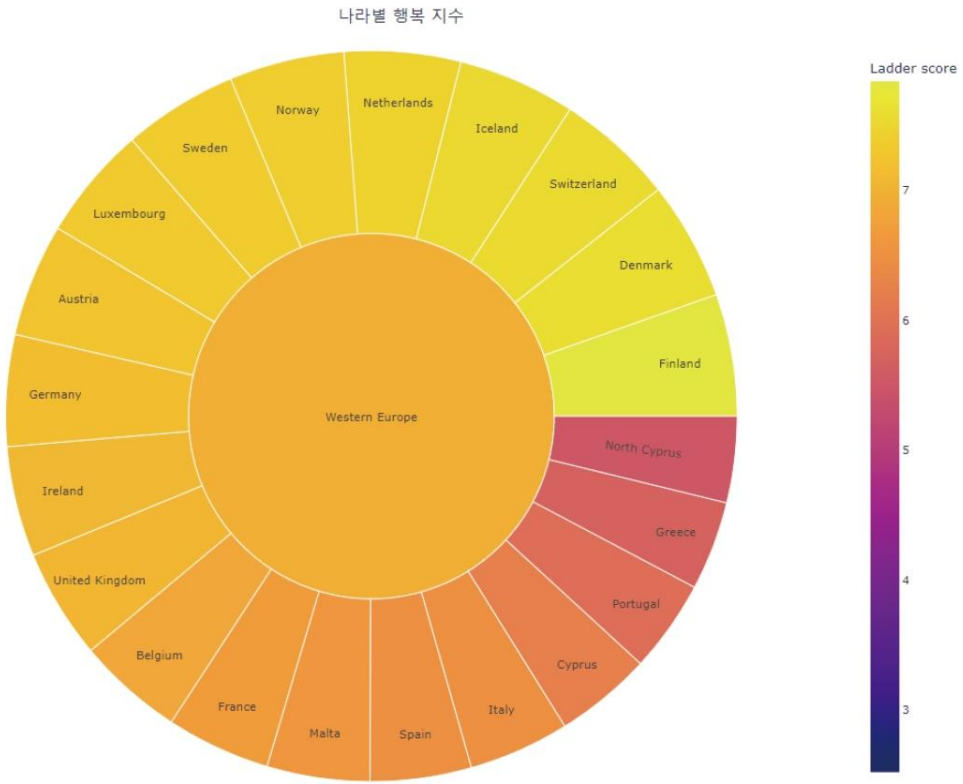


실행 결과를 통해 지역(Regional indicator)별로 속하는 나라 수 비율을 비교하기 좋았으나, 지역에 속하는 국가명 확인은 treemap이 좀 더 편리하다는 것을 확인할 수 있습니다.

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

앞 실행 결과에서 Western Europe을 클릭한 후 Finland를 선택해 보자.



실행 결과를 통해 행복 지수(Ladder score)를 기준으로 Western Europe에 속하는 국가명이 Finland부터 반시계 방향으로 오름차순 정렬되어 있는 것을 확인할 수 있습니다.

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

- **choropleth 기법 사용**

  지리 영역별 데이터 수칫값을 지도 위에 색으로 표현하는 시각화 기법을 사용한다.

```
px.choropleth(data_frame = 데이터프레임 객체,
              locations = '열 이름',
              locationmode = 'country names',
              color = '열 이름')
# locations는 열 이름(국가명)에 따라 지도에 표시
# color는 열 이름(행복 지수)에 따라 지도에 색상 표시
# locationmode는 country names 중 locations의 열 이름 항목을 일치시킴.
```

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기

locationmode의 옵션에는 ISO-3, USAstates, country names, geojson이 있지만, 이 활동에서는 국가명을 그대로 사용하는 'country names'를 사용합니다.

데이터 탐색 및 시각화하기

python

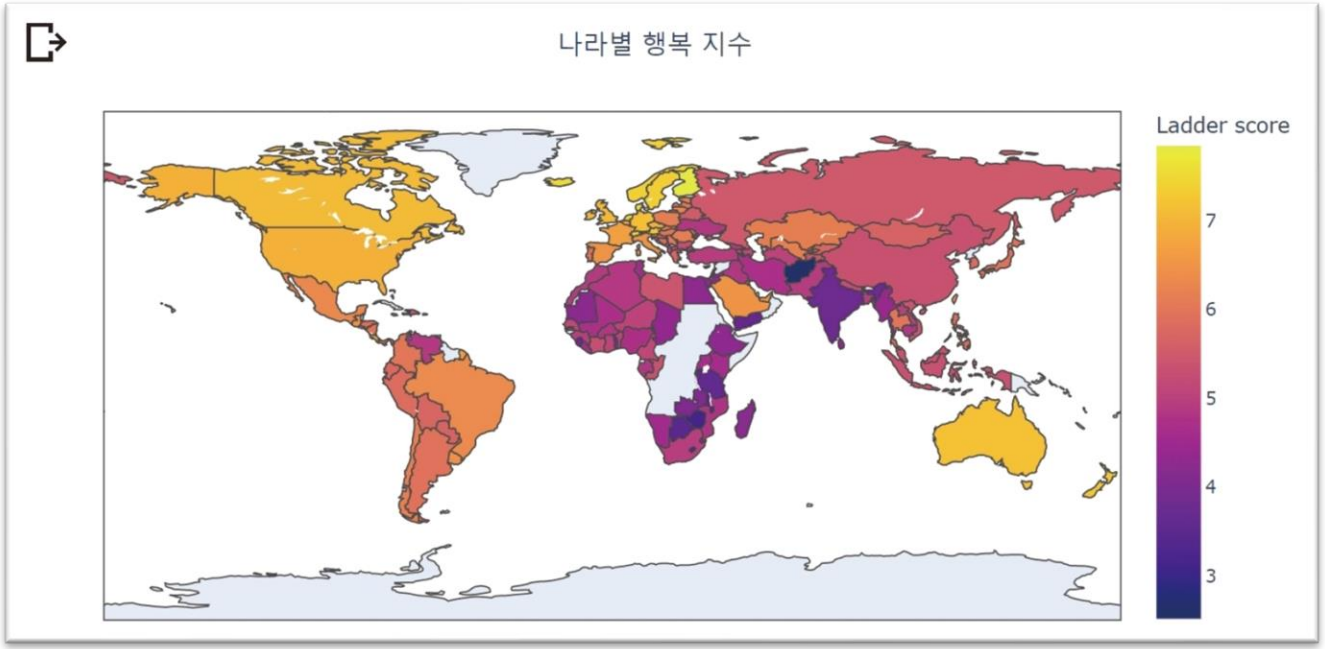
플로틀리 사용하여 시각화하기

choropleth를 사용하여 나라별 행복 지수를 지도상에 색으로 출력해보자.

```
fig = px.choropleth(data_frame = happiness,
                    locations = 'Country name',
                    locationmode = 'country names',
                    color = 'Ladder score')
fig.update_layout(title = '나라별 행복 지수', title_x = 0.5,
                  width = 900, height = 500)
fig.show
```

데이터 탐색 및 시각화하기

플로틀리 사용하여 시각화하기



locationmode의 country names와 locations의 열의 정보가 일치해야 지도에 색상이 적용됩니다.

실행 결과를 통해 색상으로 표현한 나라별 행복 지수를 지도로 한눈에 확인.

색상이 칠해지지 않은 지역은 행복 지수 산출에 참여하지 않은 국가입니다.

데이터 탐색 및 시각화하기

python

탐색정보3 알아보기

행복 지수 속성과 관련이 깊은 속성이 무엇인지 상관관계 분석을 통해 알아 보자.

상관계수 시각화하기

속성 간의 상관관계를 분석한 후 그 값을 행렬 형태로 시각화 한다.

```
px.imshow(데이터프레임 객체.corr( ), text_auto = True)
# text_auto = True: 실젯값, 상관관계(correlation) 시각화
```

데이터 탐색 및 시각화하기

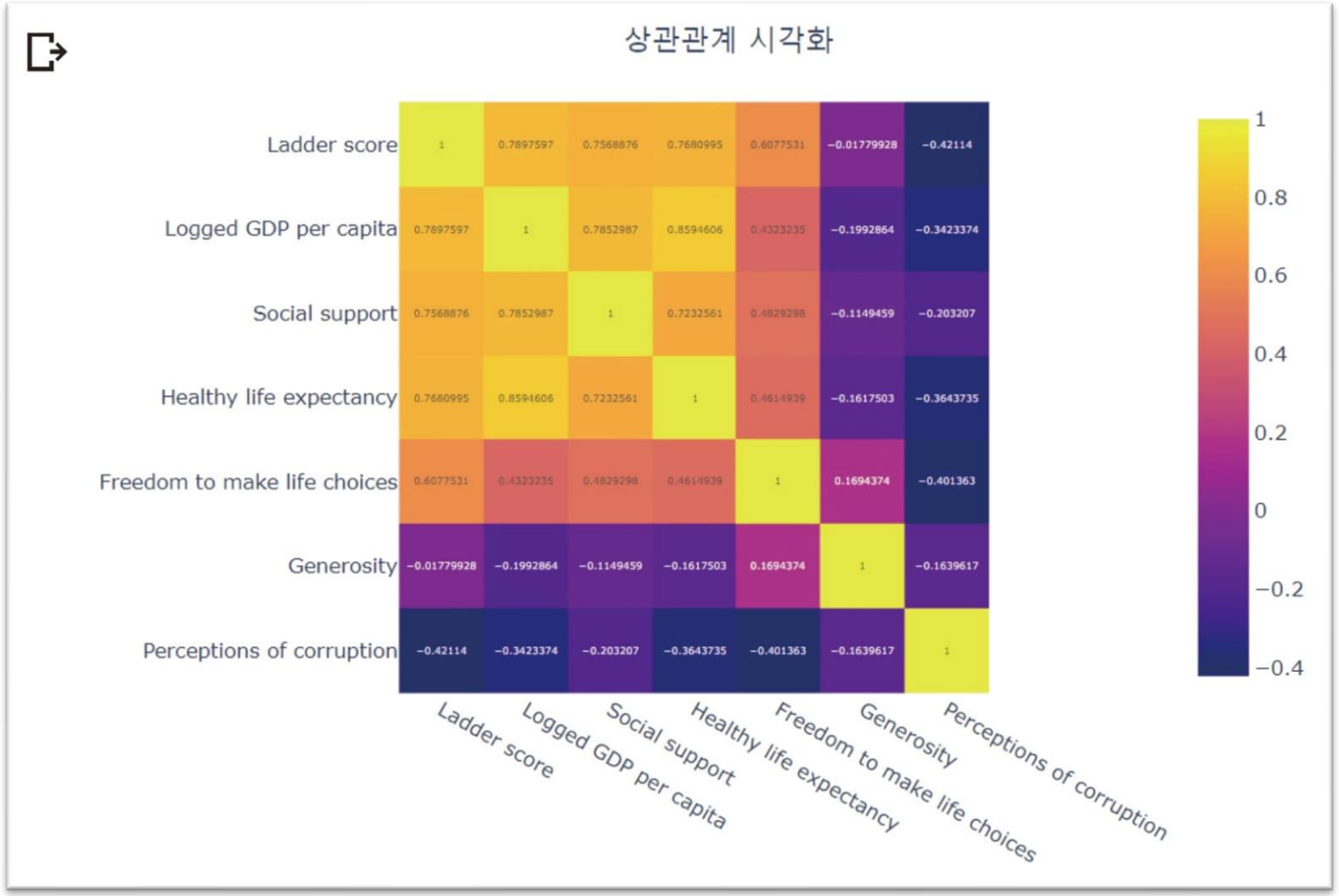
상관계수 시각화하기

행복 지수 데이터의 속성 간 상관관계를 이미지로 시각화한다.

```
1  fig = px.imshow(happiness.corr( ), text_auto = True) # 상관관계 실젯값 출력
2  fig.update_layout(title = '상관관계 시각화', width = 800, title_x = 0.5)
3  fig.show( )
```

데이터 탐색 및 시각화하기

상관계수 시각화하기



데이터 탐색 및 시각화하기

python

상관계수 시각화하기

실행 결과를 통해 각 속성 간의 상관관계를 행렬로 표현한 것을 확인할 수 있습니다. 예를 들어, 행복 지수 속성은 1인당 국내총생산(Logged GDP per capita), 사회적 지원(Social support), 건강 수명(Healthy life expectancy)과 높은 양의 상관관계를 가지며, 삶에 대한 선택의 자유(Freedom to make life choices)와는 보통의 양의 상관관계를 갖습니다. 관용(Generosity)과 부정부패 인식 지수(Perceptions of corruption)는 음의 상관관계가 있음을 확인할 수 있습니다.

데이터 탐색 및 시각화하기

산점도 사용하기

▪ 산점도 행렬(Scatter Matrix) 사용 상관관계 시각화하기

산점도 행렬(Scatter Matrix)을 사용하여 여러 개의 속성에 대하여 각 쌍을 이루도록 산점도를 그려 속성 간의 상관관계를 살펴보자.

px.scatter_matrix(데이터프레임 객체, dimensions = ['속성명'], color = '속성명')

dimensions에는 행렬로 표현할 속성명 나열,

color에는 산점도를 색으로 표현할 속성명 제시(생략 시 단색)

데이터 탐색 및 시각화하기

python

산점도 사용하기

219쪽 실행 결과를 산점도 행렬로 시각화하여 상관관계를 분석해 보자.

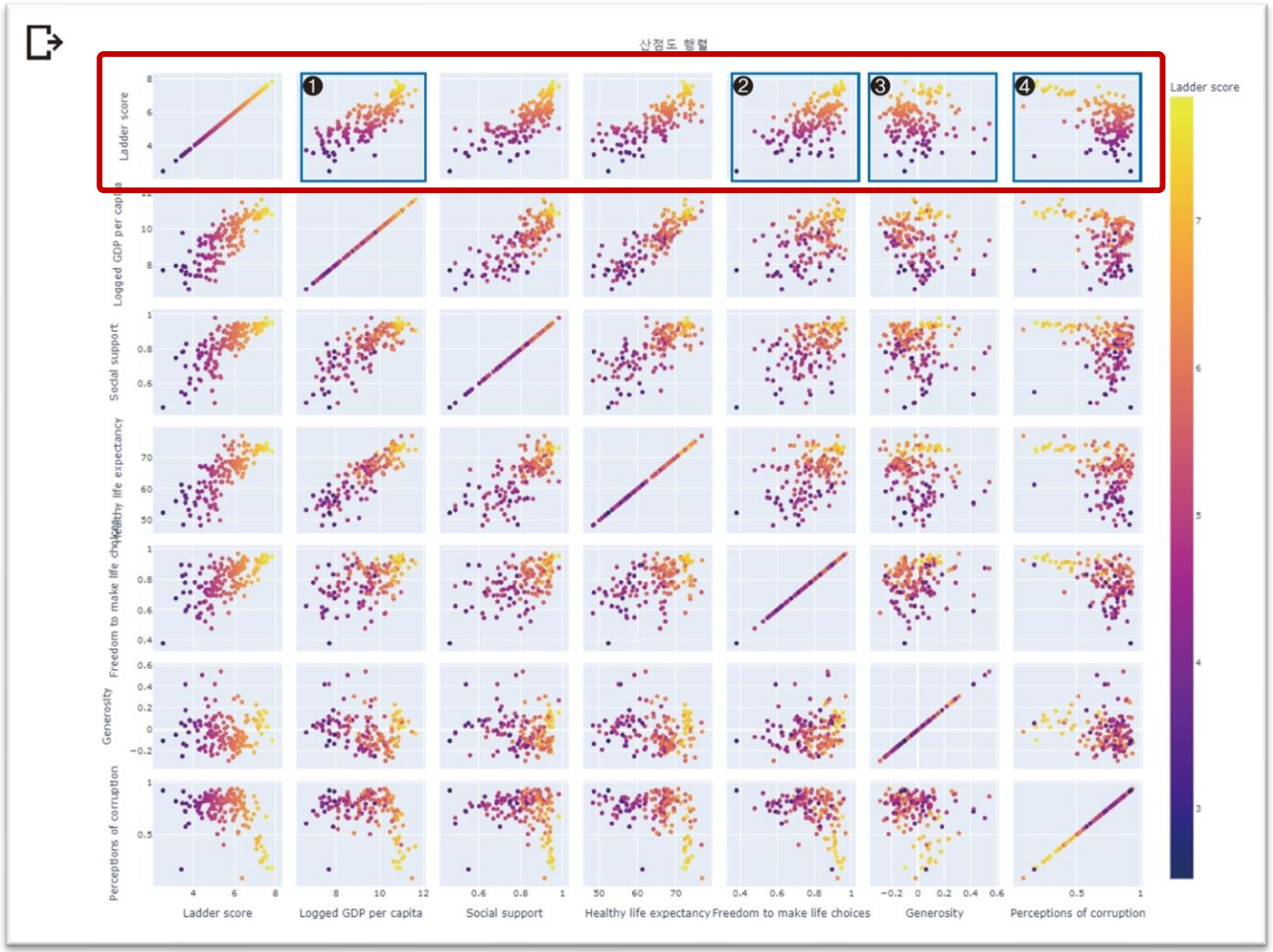
```
1 fig = px.scatter_matrix(happiness,
2         dimensions = ['Ladder score', 'Logged GDP per capita', 'Social support',
3                       'Healthy life expectancy', 'Freedom to make life choices',
4                       'Generosity', 'Perceptions of corruption'],
5                       color = 'Ladder score')
6 fig.update_layout(title = '산점도 행렬', height = 1200, title_x = 0.5)
7 fig.show
```

실행 결과는 다음 페이지에서 확인

데이터 탐색 및 시각화하기

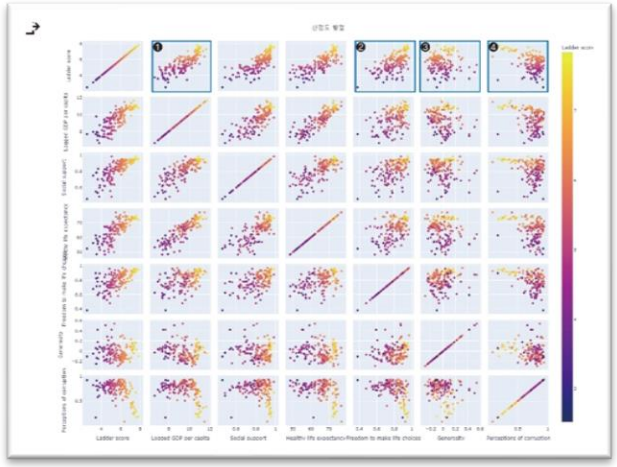
산점도 사용하기

오른쪽 실행 결과에서 표시한 부분은 행복 지수 속성과 나머지 속성과의 상관관계를 나타냅니다.



데이터 탐색 및 시각화하기

산점도 사용하기



산점도 행렬 시각화를 통해 알게 된 점

은 상관계수가 약 0.79인 높은 양의 상관관계로 속성(Logged GDP per capita)값에 따라 다른 속성(Ladder score) 값이 함께 변하는 정도가 강합니다.

는 상관계수가 약 0.61인 보통의 양의 상관관계로 속성(Freedom to make life choices) 값

에 따라 다른 속성(Ladder score) 값이 함께 변하는 정도가 보통입니다.

은 상관계수가 약 -0.02이므로 속성(Generosity) 값에 따라 다른 속성(Ladder score)값이 거의 변하지 않습니다.

는 상관계수가 약 -0.42인 보통의 음의 상관관계로 속성(Perceptions of corruption)값에 따라 다른 속성(Ladder score) 값이 함께 변하는 정도가 보통입니다.

산점도 사용하기

- **산점도(Scatter Plot)를 사용하여 상관관계 시각화하기**

산점도를 그릴 때 회귀선을 넣으면 상관관계를 이해하기 좋다.

px.scatter(데이터프레임 객체, x = 'x축 속성명', y = 'y축 속성명', size = '속성명',
trendline = '회귀선 종류', trendline_color_override = '색상')

x, y에 상관관계를 나타낼 속성명, trendline은 회귀선,
trendline_color_override는 회귀선 색상 제시

산점도 사용하기

산점도를 사용하여 1인당 국내총생산(Logged GDP per capita) 속성과 행복 지수(Ladder score) 속성과의 상관관계를 회귀선을 넣어 시각화해 보자.

```
1 fig = px.scatter(happiness, x = 'Logged GDP per capita',
2                  y = 'Ladder score', size = 'Ladder score', trendline = 'ols',
3                  trendline_color_override = 'red')
4 fig.update_layout(title = '1인당 국민총생산과 행복 지수의 상관관계',
5                  width = 800, title_x = 0.5)
6 fig.show( )
```

데이터 탐색 및 시각화하기

산점도 사용하기

관용(Generosity), 부정부패 인식 지수(Perceptions of corruption)와 행복 지수(Ladder score)의 상관관계도 산점도로 출력하여 확인해 보세요

실행 결과를 통해 1인당 국내총생산 속성이 행복 지수 속성과 관련이 깊은 것을 확인할 수 있습니다.

배운 내용 정리

문제 정의하기

문제 상황 이해하기
탐색할 정보 알아보기

데이터 불러오기

행복 지수 데이터 셋 소개
행복 지수 데이터 셋 불러오기

데이터 탐색 및 시각화하기

데이터 셋의 속성 살펴보기
중요한 속성만 선택하기(정렬, 추출 등)
다양한 그래프로 시각화하기(맷플롯립·시본·플로틀리 사용, 상관관계 분석)
시각화 결과 해석하기

배운 내용 정리

1. 우리나라의 행복 지수는 몇 위였나요? 그리고 나라별 행복 지수 순위는 어땠나요?

행복 지수 속성을 기준으로 정렬하여 확인한 결과, 우리나라의 행복 지수는 62위였습니다. 행복 지수 상위 4개 나라는 핀란드, 덴마크, 스위스, 아이슬란드이고, 행복 지수 하위 4개의 나라는 보츠나와, 르완다, 짐바브웨, 아프가니스탄이었습니다.

2. 전 세계 나라별 행복 지수를 한눈에 보기 쉽게 표현하는 방법에는 무엇이 있었나?

treemap을 통해 상위 속성과 하위 속성의 계층 구조를 시각화하여 한눈에 비교 분석할수 있었고, 부분적으로 확대하여 살펴볼 수 있었습니다.

sunburst를 통해 상위 속성과 하위 속성의 계층 구조를 마치 태양빛이 빛나는 모양으로 시각화하여 한눈에 비교 분석할 수 있었습니다.

choropleth 지도 시각화를 통해 나라별 행복 지수를 지도상에 색으로 출력하여 확인할 수 있었습니다.

배운 내용 정리

3. 행복 지수 속성과 관련 깊은 속성은 무엇이었나요?

데이터 속성 간의 상관계수를 구하고, 이를 시각화하여 살펴본 결과, 행복 지수 속성과 관련 깊은 속성은 1인당 국내총생산(Logged GDP per capita), 사회적 지원(Social support), 건강 수명(Healthy life expectancy)인 것을 확인할 수 있었습니다.