이론통계 과제 1

이화여자대학교 일반대학원 통계학과

212STG18 예지혜

Part 1. Excel Power BI 사용

- 데이터 불러오기 : 데이터 탭 – 데이터 가져오기 및 변환 – 웹 – URL 입력 – 가져올 데이터 선택 – 변환 – 원하는 모양으로 가공

[결과]

GDP 데이터

	A	B	C	D
1	Both sexes rank	Country	Continent	Both sexes
2	1	Guyana	South America	30.2
3	2	Lesotho	Africa	28.9
4	3	Russia	Europe	26.5
5	4	Lithuania	Europe	25.7
6	5	Suriname	South America	23.2
7	6	Ivory Coast	Africa	23.0
8	7	Kazakhstan	Asia	22.8
9	8	Equatorial Guinea	Africa	22.0
10	9	Belarus	Europe	21.4
11	10	South Korea	Asia	20.2

Suicide rate 데이터

	A	B	C	D
1	Both sexes rank	Country	Continent	Both sexes
2	1	Guyana	South America	30.2
3	2	Lesotho	Africa	28.9
4	3	Russia	Europe	26.5
5	4	Lithuania	Europe	25.7
6	5	Suriname	South America	23.2
7	6	Ivory Coast	Africa	23.0
8	7	Kazakhstan	Asia	22.8
9	8	Equatorial Guinea	Africa	22.0
10	9	Belarus	Europe	21.4
11	10	South Korea	Asia	20.2

행복도조사 데이터

	A	B	C
1	Overall rank	Country or region	Score
2	1	Finland	7.809
3	2	Denmark	7.646
4	3	Switzerland	7.56
5	4	Iceland	7.504
6	5	Norway	7.488
7	6	Netherlands	7.449
8	7	Sweden	7.353
9	8	New Zealand	7.3
10	9	Austria	7.294
11	10	Luxembourg	7.238

기대수명 데이터

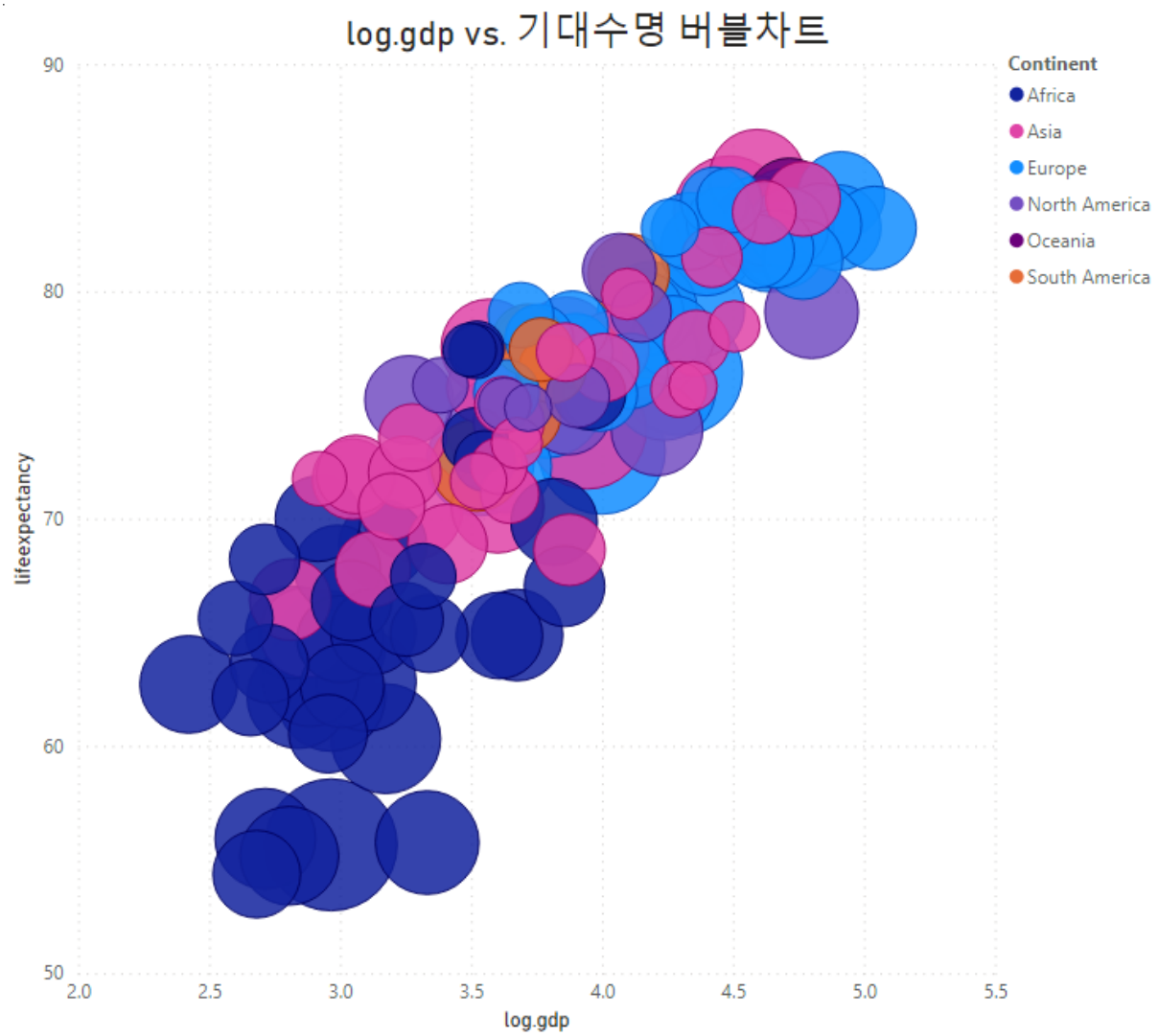
	A	B	C
1	#	Country	Life Expectancy (both sexes)
2	1	Hong Kong	85.29
3	2	Japan	85.03
4	3	Macao	84.68
5	4	Switzerland	84.25
6	5	Singapore	84.07
7	6	Italy	84.01
8	7	Spain	83.99
9	8	Australia	83.94
10	9	Channel Islands	83.6
11	10	Iceland	83.52

- 데이터 결합 : 데이터 탭 – 데이터 가져오기 – 쿼리 결합 – 병합 – 병합할 데이터 선택, 키 선택, inner join – 두 개의 데이터 씩 총 3번 결합하여 모두 결합한다

[결과]

	A	B	C	D	E	F
1	Country	Continent	gdp	suiciderate	happinessscore	lifeexpectancy
2	Albania	Europe	4898	5.6	4.883	78.96
3	Algeria	Africa	3331	3.3	5.005	77.5
4	Argentina	South America	8433	9.1	5.975	77.17
5	Armenia	Europe	4315	5.7	4.677	75.55
6	Australia	Oceania	51885	11.7	7.223	83.94
7	Austria	Europe	48634	11.4	7.294	82.05
8	Azerbaijan	Asia	4721	2.6	5.165	73.33
9	Bahrain	Asia	22878	5.7	6.227	77.73
10	Bangladesh	Asia	1888	6.1	4.833	73.57

- 버블 차트(Power BI Desktop 이용) : 데이터 탭 - log.gdp 열(=LOG('최종데이터'[gdp])) 새로 생성 - 보고서 탭 - 분산형 차트 – '자세히'는 Country, '범례'는 Continent, 'X 축'은 'log.gdp', 'Y 축'은 lifeexpectancy, '크기'는 suiciderate 으로 설정

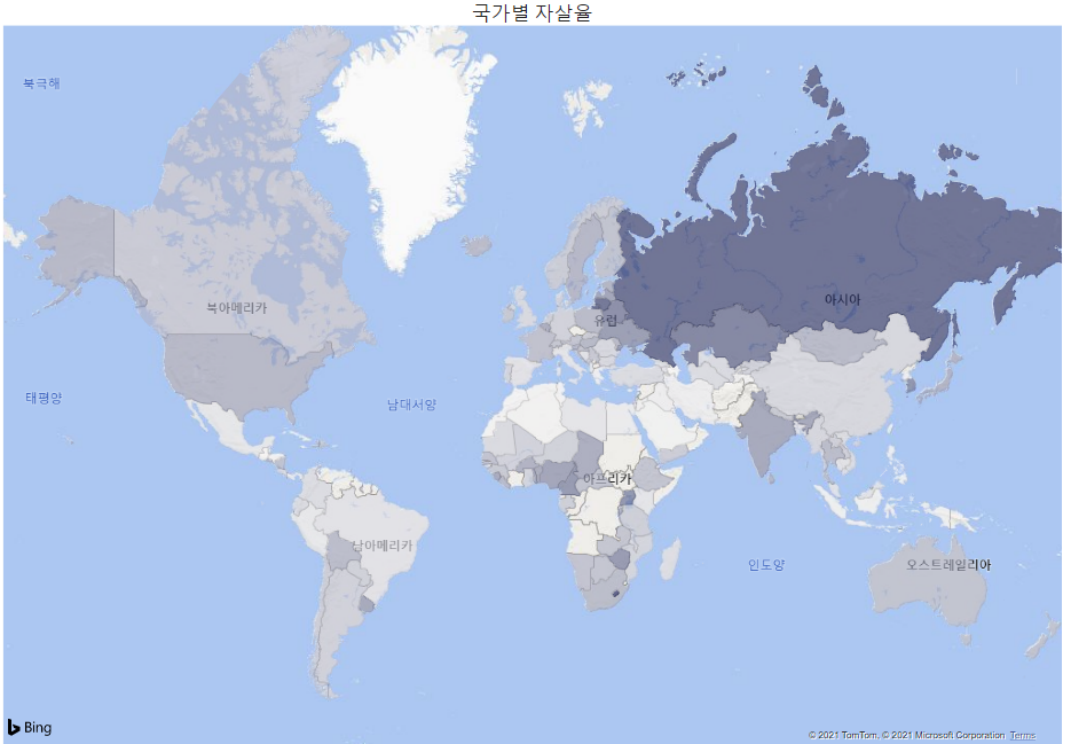


Log(gdp)와 기대수명은 양의 상관관계를 가지고 있다. 즉, gdp가 높을수록 기대수명이 높아지는 경향이 있으며, 대륙별로 살펴보면 아프리카 대륙이 낮은 gdp와 기대수명 그룹에 많이 분포해있음을 확인할 수 있다. 반대로 유럽은 gdp와 기대수명이 높은 그룹에 주로 분포해있으며, 아시아는 중반부터 높은 그룹까지 고르게 분포해있다.

크기는 자살율을 나타내는데, gdp나 기대수명과는 크게 관련성이 보이지 않는다. 경제적 또는 의료적 수준이 높다고 해서 자살율이 낮지는 않은 것이다.

- 자살율 지도 그래프(Power BI Desktop 이용) : 등치 지역도 – '위치'는 Country, '도구 설명'은 suiciderate로 설정 후 서식의 데이터색 설정

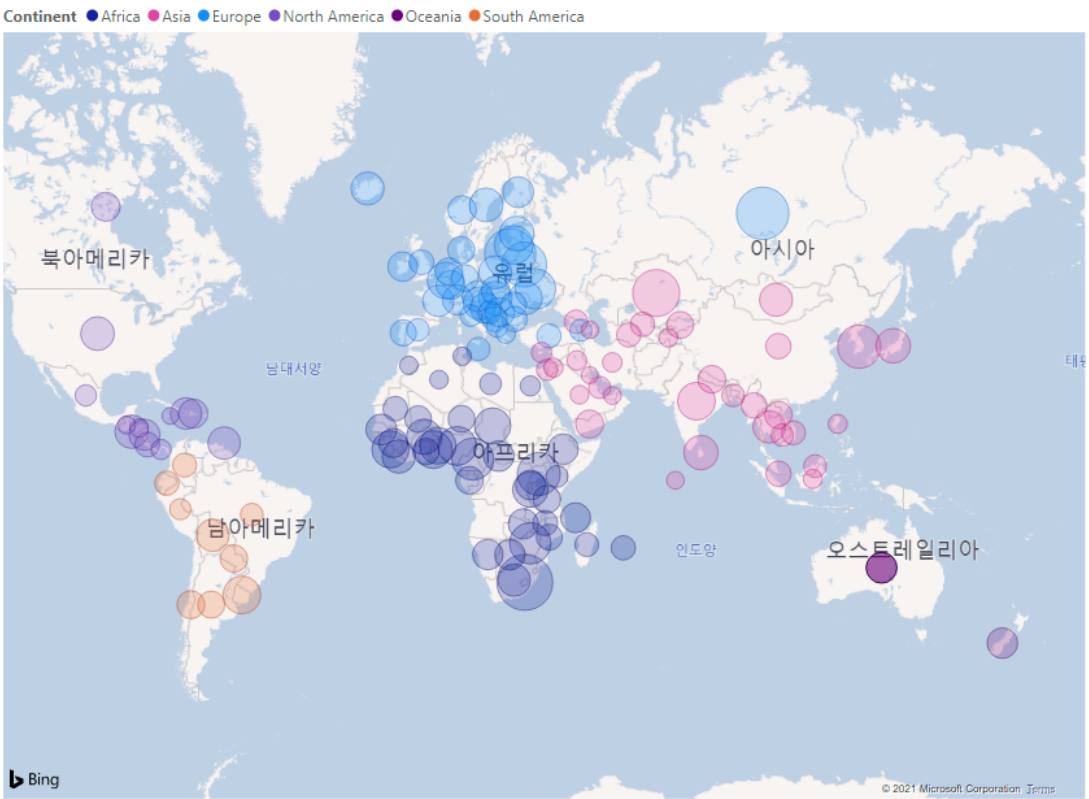
국가별 자살율



국가별 자살율을 색상으로 표현한 데이터이다. 진해질수록 자살율이 높음을 의미한다. 인도, 몽골, 러시아 등 자살율이 높은 나라가 조금 보이지만, 눈에 잘 들어오지는 않는다.

- 다른 방식 : 맵 – '위치'는 country, '범례'는 continent, '크기'는 suiciderate로 설정

국가별 자살율



Part 2. R 사용

```
library(readxl)

gdp <- read_excel("hw1.xlsx", sheet = "gdp_2020", col_names = TRUE)
gdp <- gdp[c(2,3)]
names(gdp) <- c("country","gdp")
head(gdp)
```

```
## # A tibble: 6 x 2
##   country         gdp
##   <chr>         <dbl>
## 1 Luxembourg    109602
## 2 Switzerland    81867
## 3 Ireland        79669
## 4 Norway         67989
## 5 United States  63051
## 6 Singapore      58484
```

```
suicide <- read_excel("hw1.xlsx", sheet = "suiciderate_2016", col_names =
TRUE)
suicide <- suicide[c(2,3,4)]
names(suicide) <- c("country","continent","suicide_rate")
suicide$suicide_rate <- as.double(suicide$suicide_rate)
head(suicide)
```

```
## # A tibble: 6 x 3
##   country     continent     suicide_rate
##   <chr>       <chr>               <dbl>
## 1 Guyana      South America        30.2
## 2 Lesotho     Africa              28.9
## 3 Russia      Europe              26.5
## 4 Lithuania   Europe              25.7
## 5 Suriname    South America        23.2
## 6 Ivory Coast Africa              23
```

```
happy <- read_excel("hw1.xlsx", sheet = "happinessscore_2020", col_names =
TRUE)
happy <- happy[c(2,3)]
names(happy) <- c("country","happiness_score")
head(happy)
```

```
## # A tibble: 6 x 2
##   country     happiness_score
##   <chr>                 <dbl>
## 1 Finland                7.81
## 2 Denmark                7.65
## 3 Switzerland            7.56
## 4 Iceland                7.50
## 5 Norway                 7.49
## 6 Netherlands            7.45
```

```
life <- read_excel("hw1.xlsx", sheet = "lifeexpectancy_2020", col_names =
TRUE)
```

```
life <- life[c(2,3)]
names(life) <- c("country","life_expectancy")
head(life)
```

```
## # A tibble: 6 x 2
##   country     life_expectancy
##   <chr>                 <dbl>
## 1 Hong Kong              85.3
## 2 Japan                 85.0
## 3 Macao                 84.7
## 4 Switzerland           84.2
## 5 Singapore             84.1
## 6 Italy                 84.0
```

데이터 결합

```
# merge data
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

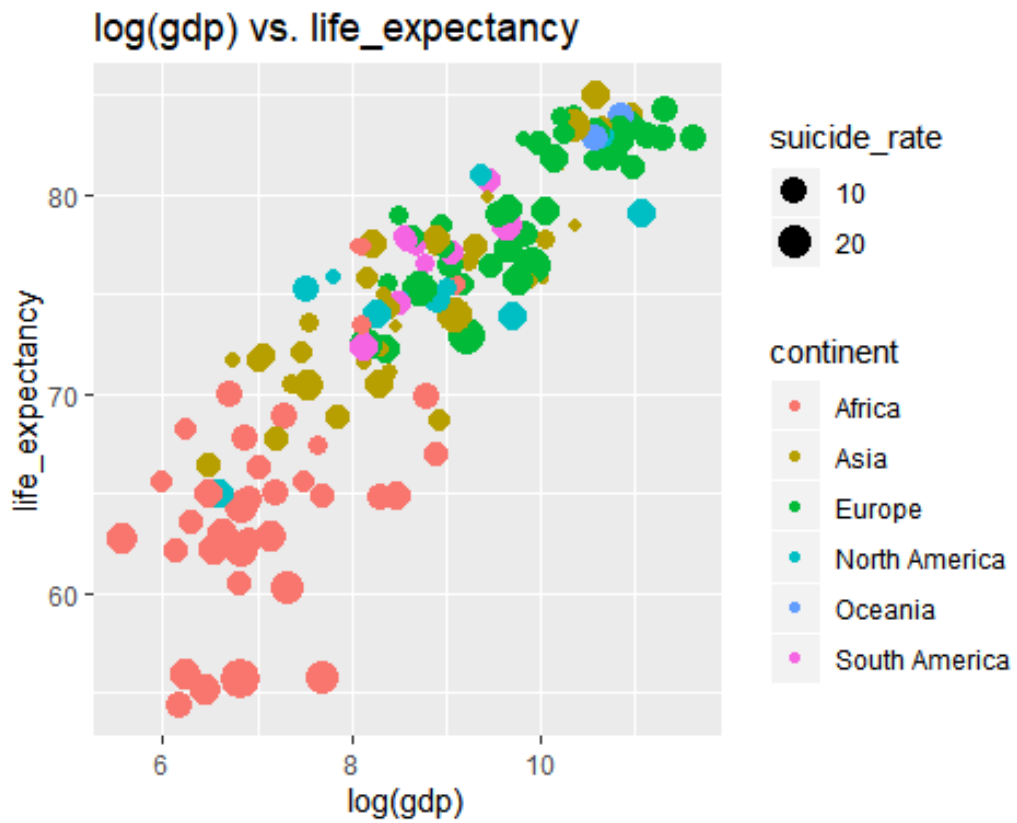
```
df <- gdp %>% inner_join(suicide, by="country") %>%
   inner_join(happy, by="country") %>% inner_join(life, by="country")
head(df)
```

```
## # A tibble: 6 x 6
##   country       gdp continent  suicide_rate happiness_score
life_expectancy
##   <chr>       <dbl> <chr>            <dbl>           <dbl>           <dbl>
## 1 Luxembourg 109602 Europe            10.4            7.24
82.8
## 2 Switzerla~  81867 Europe            11.3            7.56            84.2
## 3 Ireland     79669 Europe            10.9            7.09            82.8
## 4 Norway      67989 Europe            10.1            7.49            82.9
## 5 United St~  63051 North Ame~        13.7            6.94
79.1
## 6 Singapore   58484 Asia               7.9            6.38            84.1
```

R 의 %>% 함수를 이용하면 여러 개의 데이터도 한 번에 결합할 수 있다.

bubble chart

```
library(ggplot2)
ggplot(df, aes(x=log(gdp), life_expectancy)) +
  geom_point(aes(size=suicide_rate, color = continent)) +
  ggtitle("log(gdp) vs. life_expectancy")
```

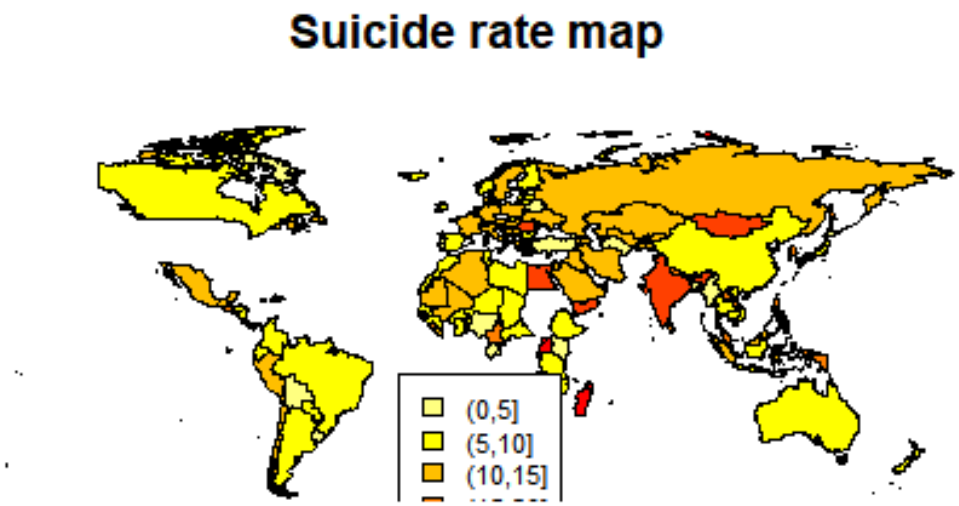


suicide rate chart

```
library(maps)

data <- data.frame(country = df$country, suicide = df$suicide_rate)
col.level <- cut(df$suicide_rate, c(0, 5, 10, 15, 20, 25, 30))
legends <- levels(col.level)
levels(col.level) <- sort(heat.colors(6), decreasing = TRUE)
data <- data.frame(data, col.level = col.level)

map('world', region = data$country, boundary = TRUE,
    fill = TRUE, col = as.character(data$col.level))
title("Suicide rate map")
legend(-25,-10, legends, fill = sort(heat.colors(6), decreasing = TRUE),
cex = 0.7)
```



Power BI는 해당되는 칸에 옵션을 넣는 반면, R은 코딩을 통해 x축과 y축, 범례 등을 설정해줘야 한다. Power BI가 훨씬 직관적이고 편리하지만, 이미 짜여진 틀 안에서만 그래프를 그릴 수 있기 때문에 확장성은 R이 더 좋다. R은 공개된 패키지를 활용할 수 있고, 많은 사람들이 이러한 패키지 제작에 참여하기 때문에 원하는 작업에 맞는 패키지가 존재한다면 편하게 이용할 수 있다. 반면, Power BI는 만들 수 있는 차트들이 직관적이고 다양하게 제시되어 있기 때문에, 이를 잘 활용한다면 같은 데이터로도 다양한 차트를 어려움없이 만들어 볼 수 있다.