|  |  |  |  |
| --- | --- | --- | --- |
| **통계패키지활용 자료분석(2024년 1학기)** | | | |
| **담 당 교 수 : 김 태 수** | | | |
| **강좌번호** | **100982-31001** | **본인의 과제**  **자체 평가** | **(7)점** |
| **과제명 :** 제3차 – 자료 시각화 실습(1) with R | | | |

|  |  |
| --- | --- |
| **이름** | **권도윤** |
|  | |

|  |  |
| --- | --- |
| **제 출 일** | **2024년 04월 05일** |
| **학 과** | **인공지능응용학과** |
| **학 번** | **21102345** |

**|목차|**

**Ⅰ. 자료 설명…………………………………………………………………………3**

**Ⅱ. 데이터 시각화….……………………………………………………………..4**

1. 산점도…………………………………………………………………………………………………………………………….4
2. 히스토그램……………………………………………………………………………………………………………………..4
3. 박스 플롯…………….………………………………………………………………………………………………………….5
4. 산점도\_2………………………………………………………………………………………………………………………….5

**Ⅲ. 결론………………………………………………………………………………….8**

**Ⅳ. 부록………………………………………………………………………………….9**

**|자료 설명|**

텍스트, 스크린샷, 폰트, 번호이(가) 표시된 사진

자동 생성된 설명본 보고서 에는 ‘국가별 행복조사의 기초자료 및 설명’ 데이터 셋을 사용했다. 이 데이터셋은 470개의 데이터와 10개의 특징들을 가지고 있으며 연도별 각 나라의 행복수치와 그와 관련된 GDP, 건강 등의 특징들을 담고있다. 다음은R프로그램의 summary 함수를 사용하여 수치적 정보를 요약한 것이다.

(국가별 행복조사의 기초자료 및 설명 데이터 summary함수 출력 결과 정리)

다음은 일부 특성을 제거한7개의 특성들 간의 상관관계를 정리한 heatmap이다. 숫자가 클수록 서로의 상관성이 크다. 지금부터 이 자료의 경향, 성격을 여러 그래프를 통해 파악해볼 것이다.

**텍스트, 스크린샷, 도표, 평행이(가) 표시된 사진

자동 생성된 설명**

(Country, Year, Happiness\_Rank 제외한 7개의 특성들에 관한 Heatmap)

**|데이터 시각화|**

**<산점도>**

**텍스트, 스크린샷, 도표, 다채로움이(가) 표시된 사진

자동 생성된 설명**

( Happiness\_Score와 다른 특징들(Country제외) 간의 관계를 나타낸 산점도)

**<히스토그램>**

텍스트, 스크린샷, 도표, 그래프이(가) 표시된 사진

자동 생성된 설명

(country를 제외한 9가지 특징들의 히스토그램)

**[설명]**

- 행복점수와 다른 특징들 사이의 상관관계를 보기 위해 출력.

**-** Year, Trust그리고 generosity를 제외하고 모두 양의 상관관계를 보임. (happiness\_rank는 그 수가 작은 것이 높은 것이기 때문에 happiness\_score와 양의 상관관계임)

**[설명]**

- 각 특징들의 분포를 보기 위해 출력.

- Year는 원래 3개의 수로 이루어져 있어 그 분포가 타당함을 알 수 있으나 Economy수치의 분포는 나머지 특징들에 비해 분포가 불균형적임. (산점도에서 Economy가 두부분으로 나뉘어 나타난 이유)

**<박스 플롯>**

도표, 스크린샷, 라인, 사각형이(가) 표시된 사진

자동 생성된 설명

(연도별 Happiness\_Score 박스 플롯)

**<산점도\_2-1>**

텍스트, 스크린샷, 번호, 라인이(가) 표시된 사진

자동 생성된 설명

(연도별 Happiness\_Score 탑5 국가)

**[설명]**

-연도별 Happiness\_Score 의 분포와 대칭성, 이상치 여부를 확인하기 위해 출력.

- 데이터의 분포가 고르고 비슷하게 이루어져 있으며 이상치가 없다.

**[설명]**

-연도별 Happiness\_Score top을 기록한 나라들을 보기 위해 출력.

- Switzerland, Icaland, Denmark, Norway, Finland 등의 나라들이 top5를 유지한 채 해당 나라들 안에서 순위가 계속 변동함.

**<산점도\_2-2>**

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

(연도별 Famliy 탑5 국가)

**<산점도\_2-3>**

텍스트, 스크린샷, 도표, 라인이(가) 표시된 사진

자동 생성된 설명

(연도별 Health(Life\_Expectancy) 탑5 국가)

**[설명]**

- Happiness\_Score수치가 높았던 나라들이 Hsppiness \_Score와 상관성이 높은 Family수치에 대해서도 높은 순위를 기록하는지 보기 위해 출력

- 2015년과 2017년 수치는 높은 Happiness\_Score를 기록한 나라들이 주로 자리하고 있지만 2016년 수치는 Norway만 그 자리를 지켰다.

**[설명]**

- Happiness\_Score수치가 높았던 나라들이 Hsppiness \_Score와 상관성이 높은 Health수치에 대해서도 높은 순위를 기록하는지 보기 위해 출력

- 3년의 수치 모두 높은 Happiness\_Score를 기록한 나라가 아닌 새로운 나라들 이 대부분을 차지하였다.

**<산점도\_2-4>**

텍스트, 스크린샷, 라인, 도표이(가) 표시된 사진

자동 생성된 설명

(연도별 Freedom 탑5 국가)

**[설명]**

- Happiness\_Score수치가 높았던 나라들이 Happiness \_Score와 상관성이 높은 Freedom수치에 대해서도 높은 순위를 기록하는지 보기 위해 출력

- 2015년과 2016년에는 높은 Happiness\_Score를 기록한 나라들이 2~3개씩 존재하지만 2016년도에는 모두 다른 국가들이 top5를 차지했다.

**|결론|**

본 과제에서 ‘국가별 행복조사의 기초자료 및 설명’ 데이터 셋으로 여러가지 정보를 시각화 해보았다. 자료는 데이터 셋의 목적을 고려하여 Happiness\_Score에 중점을 두고 그와 관련된 결과들에 대해 다른 특성들과 비교해보는 방법을 선택했다. 먼저 산점도를 보고Happiness\_Score와 양의 상관관계를 가지는 것은 Happiness\_Rank, Economy, Family, Health, Freedom인 것을 확인했다. 그러나 자료 설명 부분에 첨부했던 히트맵을 보면 Economy와 Happiness\_Score의 상관관계 수치가 0.27로 낮게 나오는 것을 볼 수 있다. 그 이유는 히스토그램에서 Economy 데이터의 분포가 불균형 하게 이루어져 있는 것을 통해 알 수 있다. 그리고 앞서 자료 해석에 사용되지 않았던 Year, Country 특성과 Happiness\_Score 간의 관계 정보를 얻고자 먼저 연도별 Happiness\_Score에대해 박스 플롯 형태로 출력했고 Happiness\_Score가 연도에 관계없이 비슷한 분포를 가지는 것을 확인했다. 또한 Country 특성을 알아 보기위해 Happiness\_ Score 중 상위 다섯개의 국가를 각 연도에 따라 산점도로 표현했고 그것을 통해 Switzerland, Icaland, Denmark, Norway, Finland 5개의 국가가 계속 높은 수치의 Happiness\_Score를 기록하는 것을 알 수 있었다.

Happiness\_Score와 상관성이 큰 세가지 특성들Family, Health, Freedom 역시 특정 5개국이 높은 수치를 기록할 것 이라는 예측을 가지고 데이터를 시각화 해보았다. 예상과 달리 타국가들이 top5 에 섞여 있었다. 따라서 행복지수는 상관성이 높은 특성들에 의해 정해지는 것이 아닌 여러가지 특성을 전부 고려해야 한다는 결론이 도출되었다. 특성에 대한 중요도는 하나하나 시각화 해봄으로 파악하는 것은 어렵고 시간이 오래 걸릴 것이다. 이 경우 머신러닝을 사용하여 가중치를 학습해본다면 어떤 특성들이 행복지수과 밀접한 관계를 가지는지 자세히 알 수 있을 것이다.

**|부록|**

**- 사용한 R스크립트**

library(openxlsx)

# 엑셀 데이터 가져오기

data <- read.xlsx("C:/국가별 행복조사의 기초자료및설명.xlsx")

# 수치화된 요약 출력

summary(data)

library(gplots)

library(dplyr)

data\_df <- as.data.frame(data)

# 히트맵 생성

library(reshape2)

data\_without\_some <- data\_df |>

select(-Country, -Year, -Happiness.Rank)

correlation\_matrix <- cor(data\_without\_some)

cor\_df <- melt(correlation\_matrix)

heatmap\_plot <- ggplot(data = cor\_df, aes(x = Var1, y = Var2, fill = value,label = round(value, 2))) +

geom\_tile() +

geom\_text(color = "black", size = 3) +

scale\_fill\_gradient(low = "white", high = "red") +

labs(title = "Heatmap of Features", x = "Features", y = "Features") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 90, vjust = 0.5, hjust = 1))

print(heatmap\_plot)

# Happiness\_Score와 다른 특성들 산점도

library(ggplot2)

library(gridExtra)

grid.arrange(

qplot(Happiness.Score, Year, data = data\_df, geom = "point", color = I("green")),

qplot(Happiness.Score, Happiness.Rank, data = data\_df, geom = "point", color = I("red")),

qplot(Happiness.Score, `Economy.(GDP.per.Capita)`, data = data\_df, geom = "point", color = I("purple")),

qplot(Happiness.Score, Family, data = data\_df, geom = "point", color = I("orange")),

qplot(Happiness.Score, `Health.(Life.Expectancy)`, data = data\_df, geom = "point", color = I("cyan")),

qplot(Happiness.Score, Freedom, data = data\_df, geom = "point", color = I("yellow")),

qplot(Happiness.Score, `Trust.(Government.Corruption)`, data = data\_df, geom = "point", color = I("magenta")),

qplot(Happiness.Score, Generosity, data = data\_df, geom = "point", color = I("black")),

nrow = 3

)

# 각 특징들의 분포 히스토그램

grid.arrange(

qplot(Year, data= data\_df, geom = "histogram", color = I("green")),

qplot(Happiness.Rank, data = data\_df, geom = "histogram", color = I("red")),

qplot(Happiness.Score, data = data\_df, geom = "histogram", color = I("gray")),

qplot(`Economy.(GDP.per.Capita)`, data = data\_df, geom = "histogram", color = I("purple")),

qplot(Family, data = data\_df, geom = "histogram", color = I("orange")),

qplot(`Health.(Life.Expectancy)`, data = data\_df, geom = "histogram", color = I("cyan")),

qplot(Freedom, data = data\_df, geom = "histogram", color = I("yellow")),

qplot(`Trust.(Government.Corruption)`, data = data\_df, geom = "histogram", color = I("magenta")),

qplot(Generosity, data = data\_df, geom = "histogram", color = I("black")),

nrow = 3

)

# 연도별 행복점수 박스플롯

data\_df$Year <- factor(data\_df$Year)

qplot(Year, Happiness.Score, data =

data\_df, geom = "boxplot", fill = Year) +

geom\_rug() +

labs(x = "Year", y = "Happiness Score") +

theme\_minimal()

# 연도별 행복 순위 탑5 나라 뽑기

selected\_data <- data\_df |>

select(Year, Country, Happiness.Score)

selected\_data\_2015 <- selected\_data |>

filter(Year == 2015)

top5\_2015 <- selected\_data\_2015 |>

slice\_max(n = 5, order\_by = Happiness.Score)

selected\_data\_2016 <- selected\_data |>

filter(Year == 2016)

top5\_2016 <- selected\_data\_2016 |>

slice\_max(n = 5, order\_by = Happiness.Score)

selected\_data\_2017 <- selected\_data |>

filter(Year == 2017)

top5\_2017 <- selected\_data\_2017 |>

slice\_max(n = 5, order\_by = Happiness.Score)

combined\_df <- rbind(top5\_2015, top5\_2016, top5\_2017)

ggplot(data = combined\_df, aes(x = Year, y = Happiness.Score, label = Country)) +

geom\_point(color = "red") +

geom\_text(vjust = -0.5) +

labs(x = "Year", y = "Happiness\_Score", title = "연도별 행복수치 top\_5 국가") +

theme\_minimal()+

theme(plot.title = element\_text(hjust = 0.5, margin = margin(t = 20)))

# 연도별 family 상위 5개국 뽑기

select\_Family <- data\_df |>

select(Year, Country, Family)

select\_Family\_2015 <- select\_Family |>

filter(Year == 2015)

top5\_Family\_2015 <- select\_Family\_2015 |>

slice\_max(n = 5, order\_by = Family)

top5\_Family\_2015

select\_Family <- data\_df |>

select(Year, Country, Family)

select\_Family\_2016 <- select\_Family |>

filter(Year == 2016)

top5\_Family\_2016 <- select\_Family\_2016 |>

slice\_max(n = 5, order\_by = Family)

top5\_Family\_2016

select\_Family <- data\_df |>

select(Year, Country, Family)

select\_Family\_2017 <- select\_Family |>

filter(Year == 2017)

top5\_Family\_2017 <- select\_Family\_2017 |>

slice\_max(n = 5, order\_by = Family)

top5\_Family\_2017

combined\_df <- rbind(top5\_Family\_2015, top5\_Family\_2016, top5\_Family\_2017)

library(ggrepel)

ggplot(data = combined\_df, aes(x = Year, y = Family, label = Country)) +

geom\_point(color = "red") +

geom\_text\_repel() +

labs(x = "Year", y = "Family", title = "연도별 Family수치 top\_5 국가") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5, margin = margin(t = 20)))

# 연도별 heath 상위 5개국 뽑기

select\_Health <- data\_df |>

select(Year, Country, `Health.(Life.Expectancy)`)

select\_Health\_2015 <- select\_Health |>

filter(Year == 2015)

top5\_Health\_2015 <- select\_Health\_2015 |>

slice\_max(n = 5, order\_by = `Health.(Life.Expectancy)`)

top5\_Health\_2015

select\_Health <- data\_df |>

select(Year, Country, `Health.(Life.Expectancy)`)

select\_Health\_2016 <- select\_Health |>

filter(Year == 2016)

top5\_Health\_2016 <- select\_Health\_2016 |>

slice\_max(n = 5, order\_by = `Health.(Life.Expectancy)`)

top5\_Family\_2016

select\_Health <- data\_df |>

select(Year, Country, `Health.(Life.Expectancy)`)

select\_Health\_2017 <- select\_Health |>

filter(Year == 2017)

top5\_Health\_2017 <- select\_Health\_2017 |>

slice\_max(n = 5, order\_by = `Health.(Life.Expectancy)`)

top5\_Health\_2017

combined\_df <- rbind(top5\_Health\_2015, top5\_Health\_2016, top5\_Health\_2017)

ggplot(data = combined\_df, aes(x = Year, y = `Health.(Life.Expectancy)`, label = Country)) +

geom\_point(color = "red") +

geom\_text\_repel() +

labs(x = "Year", y = "Health(Life\_Expectancy)", title = "연도별 Health수치 top\_5 국가") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5, margin = margin(t = 20)))

# 연도별 freedom 상위 5개국 뽑기

select\_Freedom <- data\_df |>

select(Year, Country, Freedom)

select\_Freedom\_2015 <- select\_Freedom |>

filter(Year == 2015)

top5\_Freedom\_2015 <- select\_Freedom\_2015 |>

slice\_max(n = 5, order\_by = Freedom)

top5\_Freedom\_2015

select\_Freedom <- data\_df |>

select(Year, Country, Freedom)

select\_Freedom\_2016 <- select\_Freedom |>

filter(Year == 2016)

top5\_Freedom\_2016 <- select\_Freedom\_2016 |>

slice\_max(n = 5, order\_by = Freedom)

top5\_Freedom\_2016

select\_Freedom <- data\_df |>

select(Year, Country, Freedom)

select\_Freedom\_2017 <- select\_Freedom |>

filter(Year == 2017)

top5\_Freedom\_2017 <- select\_Freedom\_2017 |>

slice\_max(n = 5, order\_by = Freedom)

top5\_Freedom\_2017

combined\_df <- rbind(top5\_Freedom\_2015, top5\_Freedom\_2016, top5\_Freedom\_2017)

ggplot(data = combined\_df, aes(x = Year, y = Freedom, label = Country)) +

geom\_point(color = "red") +

geom\_text\_repel() +

labs(x = "Year", y = "Freedom", title = "연도별 Freedom수치 top\_5 국가") +

theme\_minimal() +

theme(plot.title = element\_text(hjust = 0.5, margin = margin(t = 20)))