R 과제

고명섭

2020-06-18

library(ggplot2)  
library(tidyverse)

## -- Attaching packages ----------------------------------------------------------------------------------------------------- tidyverse 1.3.0 --

## √ tibble 3.0.1 √ dplyr 1.0.0  
## √ tidyr 1.1.0 √ stringr 1.4.0  
## √ readr 1.3.1 √ forcats 0.5.0  
## √ purrr 0.3.4

## -- Conflicts -------------------------------------------------------------------------------------------------------- tidyverse\_conflicts() --  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(patchwork)  
  
options("scipen" = 100)

# 1. R 스튜디오를 이용하여 다음 작업을 수행하시오.

### 1) 작업폴더 ’c:data/R\_exam’을 생성하고 이를 기본경로로 지정하라.

# setwd("c:data/R\_exam")

### 2) R에 내장되어 있는 Orange데이터의 기술통계량을 출력하라.

summary(Orange)

## Tree age circumference   
## 3:7 Min. : 118.0 Min. : 30.0   
## 1:7 1st Qu.: 484.0 1st Qu.: 65.5   
## 5:7 Median :1004.0 Median :115.0   
## 2:7 Mean : 922.1 Mean :115.9   
## 4:7 3rd Qu.:1372.0 3rd Qu.:161.5   
## Max. :1582.0 Max. :214.0

### 3) 위의 작업을 수행했던 작업공간을 기본 경로에 저장하라.

# 2. 다음과 같은 작업을 위한 R 스크립트를 작성하고 결과를 출력하시오

### 1) 1에서 10까지의 정수 중 5개를 랜덤하게 추출하여 백터 x를 생성하라

x <- sample(1:10,5)

### 2) 2에서 11까지의 정수중 5개를 랜덤하게 추출하여 벡터 y를 생성하라.

y <- sample(2:11, 5)

### 3) x원소의 합과 y원소의 합을 비교 하여 결과를 출력하는 명령문을 작성하라

if(sum(x) == sum(y)) {  
 print("같음")  
} else if (sum(x) > sum(y)) {  
 print("x의 합이 더 큼")  
} else {  
 print("y의 합이 더큼")  
}

## [1] "x의 합이 더 큼"

# 3. 구구단을 출력하는 프로그램을 작성하고자 한다 , gugu라는 변수에 초기값이 0인 9x9행렬을 만들고 while문으로 이루어진 중첩루프를 이용하여 아래와 같은 결과를 얻을 수 있도록 R 스크립트를 작성하시오.

gugu <- matrix(nrow = 9, ncol = 9, rep(0,9\*9))  
  
i <- 1  
  
while(i <= 9) {  
 print(i)  
 gugu[i,] <- i \* c(1:9)  
 i <- i +1  
}

## [1] 1  
## [1] 2  
## [1] 3  
## [1] 4  
## [1] 5  
## [1] 6  
## [1] 7  
## [1] 8  
## [1] 9

gugu

## [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9]  
## [1,] 1 2 3 4 5 6 7 8 9  
## [2,] 2 4 6 8 10 12 14 16 18  
## [3,] 3 6 9 12 15 18 21 24 27  
## [4,] 4 8 12 16 20 24 28 32 36  
## [5,] 5 10 15 20 25 30 35 40 45  
## [6,] 6 12 18 24 30 36 42 48 54  
## [7,] 7 14 21 28 35 42 49 56 63  
## [8,] 8 16 24 32 40 48 56 64 72  
## [9,] 9 18 27 36 45 54 63 72 81

# 4. 지난 한 주간의 방문지(직장, 집, 학교 등)에 대해 이동교통수단, 이동시간, 방문지에 머문시간, 소비 지출 등의 가상 데이터를 작성하고 이를 그림으로 표현하는 네트워크 그림을 작성하라.

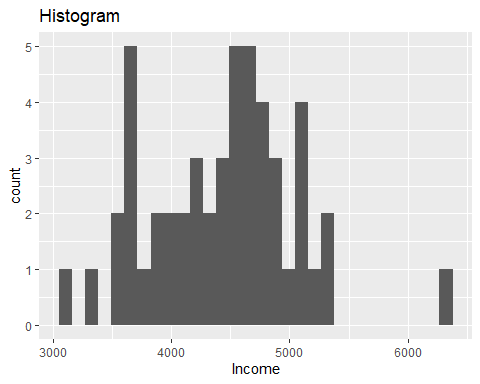
# 5. R의 기본패키지에 내장되어 있는 state.x77데이터의 소득(income) 변수에 대한 히스토그램 및 확률밀도함수 그림을 그려보라. 또한 Q-Q plot을 통해 자료가 정규분포를 따르는지 확인해보라.

* histogram: <https://rfriend.tistory.com/67>
* Normal Distribution: <https://rfriend.tistory.com/95>
* Q-Q plot: <https://ggplot2.tidyverse.org/reference/geom_qq.html>
* ggplot: <http://blog.naver.com/PostView.nhn?blogId=nife0719&logNo=221000580841>

### 1. histogram

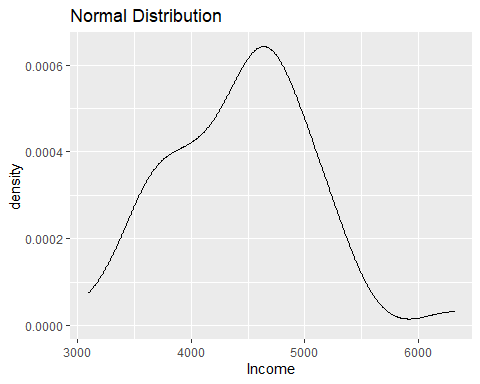
state.x77 %>%   
 as.data.frame() %>%   
 ggplot(aes(x=Income)) +   
 geom\_histogram() +  
 labs(title = "Histogram")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



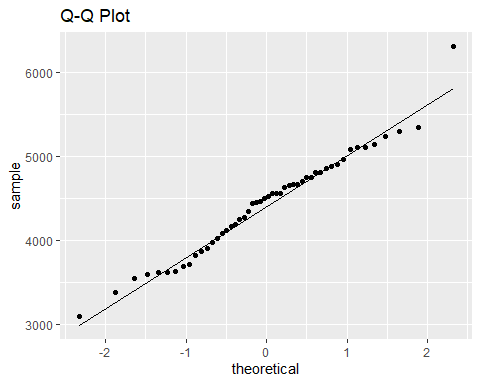
### 2. 확률밀도 함수

state.x77 %>%   
 as.data.frame() %>%   
 ggplot(aes(x=Income)) +  
 geom\_density() +   
 labs(title = "Normal Distribution")



### 3. Q-Q plot

state.x77 %>%   
 as.data.frame() %>%   
 ggplot(aes(sample = Income)) +  
 stat\_qq() +  
 stat\_qq\_line() +  
 labs(title = "Q-Q Plot")

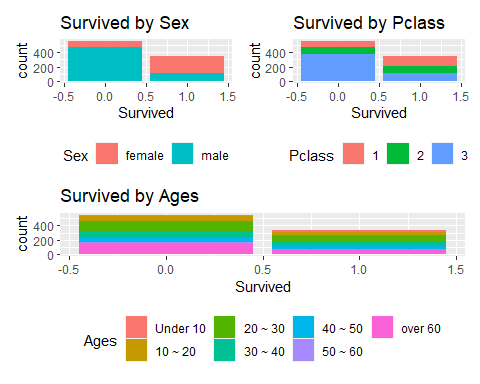


# 1. 런던 콜레라지도를 조사하고 데이터 시각화의 의미를 정리하라

# 2. Titanic 자료를 바탕으로 사고시 생존을 결정하는 중요한 요인을 시각화를 통해 찾으라.

출처: “<https://unfinishedgod.github.io/docs/kaggle/Titanic_Machine_Learning_from_Disaster/titanic.html>”

df <- read.csv("train.csv")  
  
df$Pclass <- as.factor(df$Pclass)  
  
df <- df %>%   
 mutate(Ages = case\_when(  
 Age < 10 ~ "Under 10",  
 Age < 20 ~ "10 ~ 20",  
 Age < 30 ~ "20 ~ 30",  
 Age < 40 ~ "30 ~ 40",  
 Age < 50 ~ "40 ~ 50",  
 Age < 60 ~ "50 ~ 60",  
 TRUE ~ "over 60"  
 ))   
  
df$Ages <-   
 factor(df$Ages,  
 levels = c("Under 10", "10 ~ 20", "20 ~ 30", "30 ~ 40", "40 ~ 50", "50 ~ 60", "over 60"))  
  
plot\_1 <- ggplot(df, aes(x=Survived, fill = Sex)) +  
 geom\_bar() +  
 ggtitle("Survived by Sex") +  
 theme(legend.position="bottom")  
  
plot\_2 <- ggplot(df, aes(x = Survived, fill = Pclass)) +  
 geom\_bar() +  
 ggtitle("Survived by Pclass") +  
 theme(legend.position="bottom")   
  
plot\_3 <- df %>%   
 ggplot(aes(x = Survived, fill = Ages)) +  
 geom\_bar() +  
 ggtitle("Survived by Ages") +  
 theme(legend.position="bottom")   
  
(plot\_1 + plot\_2) / plot\_3



1. 2010년, 2019년 우리나라 지역별 연간 실업률을 지도 상에 표현하는 지리적 데이터 시각화를 작성하고 2010년의 지역별 실업률과 2019년의 지역별 실업률을 비교하라.(연간 실업률 데이터는 KOSIS-노동)
2. 아름답고 의미 있는 데이터 시각화 사례를 발굴하고 그 이유를 정리 하시오.

tidycovid19패키지의 covid19 시각화. 코로나19의 전세계적으로 방대한 데이터를 정리 했을 뿐만 아니라, 코로나발생 현황의 지도를 표현 하였다. <https://github.com/joachim-gassen/tidycovid19>