R

[R] 같은 형태의 ggplot 과 barplot 만들기 (차이 비교)

2020. 7. 30. 22:15 수정 삭제 공개

ggplot과 barplot이 헷갈리는 경우가 있어 한번 정리했다.

```
> #범주형 vs 범주형 가지고 데이터 분포를 확인한다면?
> # 1. resident2 , age2 를 범주형으로 변환
>
> dataset$resident2 ← factor(dataset$resident2)
> dataset$age2 ← factor(dataset$age2)
>
> str(dataset)
'data.frame':   231 obs. of  15 variables:
 $ resident : int  1 2 4 5 3 2 2 5 3 1 ...
 $ gender   : int  1 1 2 1 1 2 1 2 1 1 ...
 $ job      : int  1 2 NA 3 2 1 2 NA 3 1 ...
 $ age      : int  26 54 45 62 57 36 37 29 35 56 ...
 $ position : int  4 1 2 1 NA 3 3 4 4 1 ...
 $ price    : num  5.1 4.2 3.5 5 5.4 4.1 4.9 2.3 4.2 6.7 ...
 $ survey   : int  5 4 4 5 4 2 3 1 3 4 ...
 $ price2   : num  5.1 4.2 3.5 5 5.4 4.1 4.9 2.3 4.2 6.7 ...
 $ price3   : num  5.1 4.2 3.5 5 5.4 4.1 4.9 2.3 4.2 6.7 ...
 $ resident2: Factor w/ 5 levels "1.서울특별시",..: 1 2 4 5 3 2 2 5 3 1 ...
 $ job2     : chr  "공무원" "회사원" NA "개인사업" ...
 $ age2     : Factor w/ 3 levels "장년층","중년층",..: 3 2 2 1 1 2 2 3 2 1 ...
 $ position2: chr  "4급" "1급" "2급" "1급" ...
 $ gender2  : chr  "남자" "남자" "여자" "남자" ...
 $ age3     : int  1 2 2 3 3 2 2 1 2 3 ...
> levels(dataset$resident2)
[1] "1.서울특별시" "2.인천광역시" "3.대전광역시" "4.대구광역시" "5.시구군"
> levels(dataset$age2)
[1] "장년층" "중년층" "청년층"
```

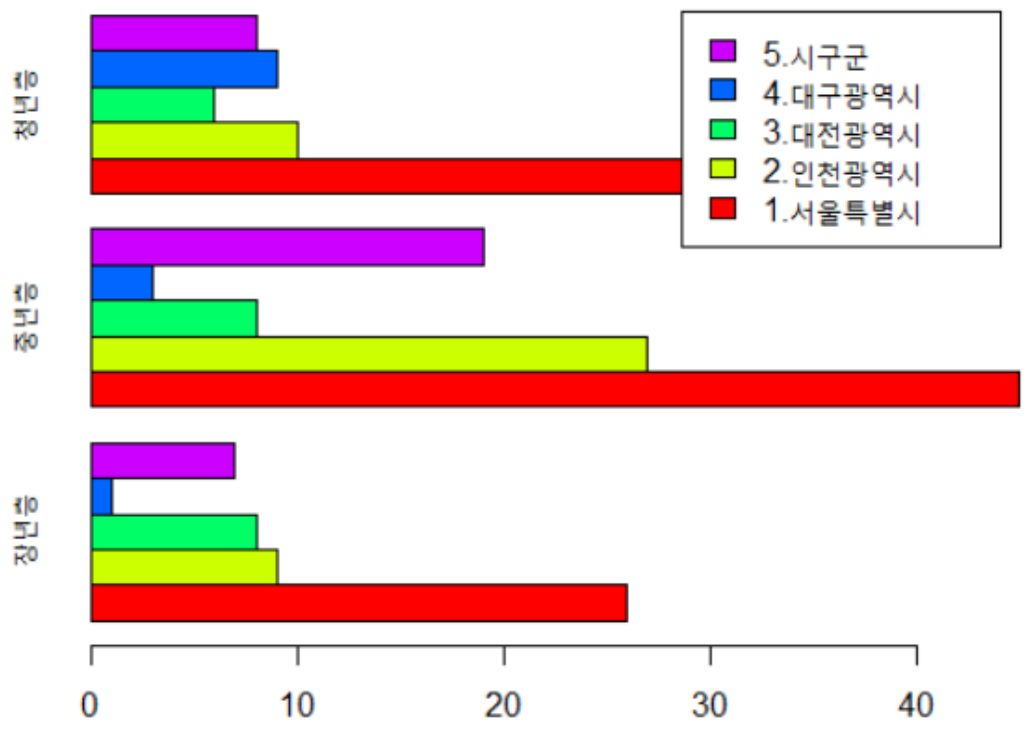
```
> # 2. 두 변수를 table()이용하여 분포를 확인해보자
> resident_gender<-table(dataset$resident2,dataset$age2)
> resident_gender

             장년층 중년층 청년층
  1.서울특별시     26     45     31
  2.인천광역시      9     27     10
  3.대전광역시      8      8      6
  4.대구광역시      1      3      9
  5.시구군          7     19      8
> class(resident_gender)
[1] "table"
>
```

barplot

데이터는 백터 또는 행렬로 받으면 된다.

```
barplot(resident_gender,
      horiz =T,         # 그래프를 90도 회전한다.
      beside = T,       # TRUE를 지정하면 그룹을 묶어서 각각의 값마다 막대를 그린다.
      legend = row.names(resident_gender), # 범례
      col = rainbow(5))              # 색깔을 무지개색깔중 5개를 골라서한다.
```

```
28  ## 같은 형태의 barplot 과 ggplot 만들기
29
30  #barplot
31  ?barplot
32  barplot(resident_gender,
33          horiz =T,
34          beside = T,
35          legend = row.names(resident_gender),
36          col = rainbow(5))
37
```

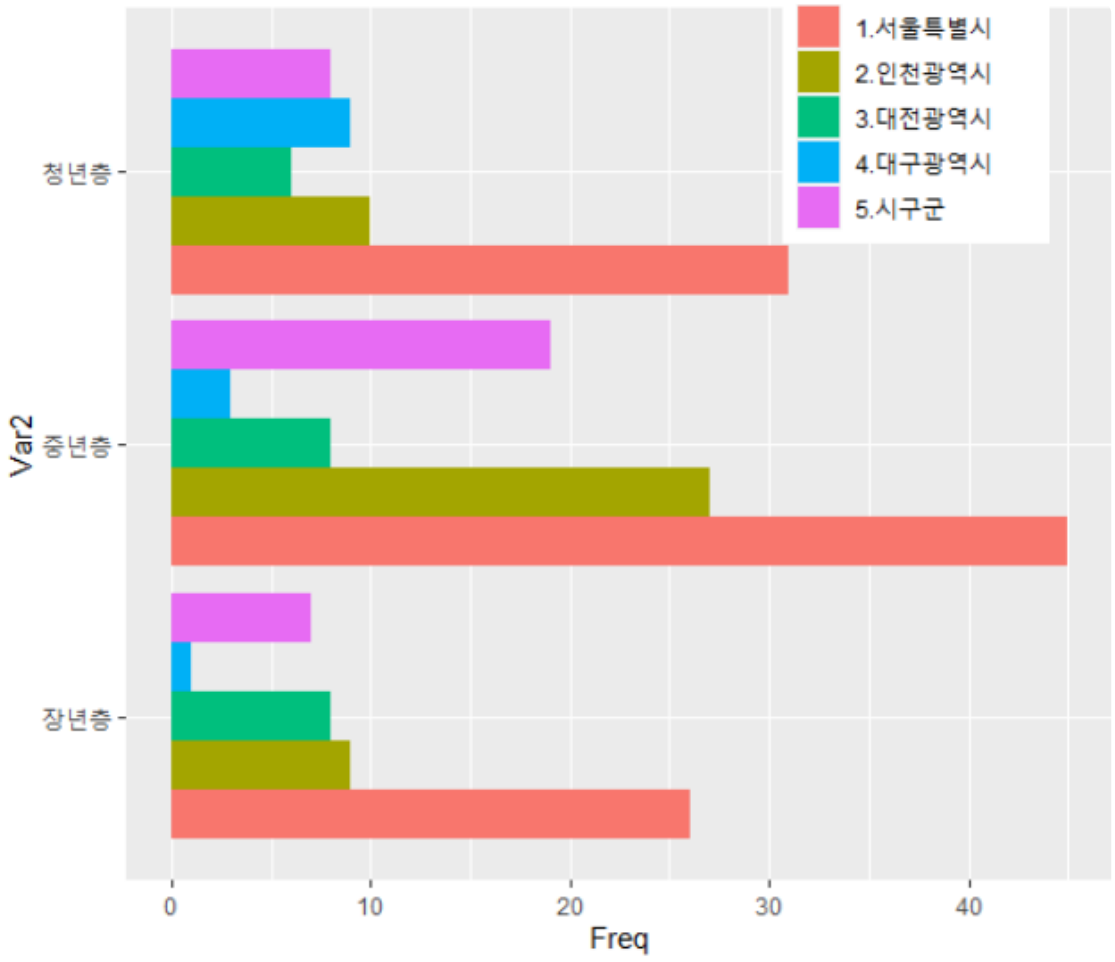


#ggplot

데이터를 넣을 때 반드시 데이터 프레임으로 받아야 된다.

```
#ggplot을 통해 처음 그래프 틀을 만든다. 거기에 x축을 빈도 y축을 연령층으로 잡고 색깔 구분을 지
ggplot(resident_gender_df , aes(x=Freq, y=Var2, fill=Var1 ) )+
 geom_bar(stat = "identity",position='dodge')+ # 그 후 geom_bar 명령어를 써서 막대그래프를 그리는
 theme(legend.position = c(.8, .90))          # 기존에 하나의 축만가능해서 stat='identity'를 사용해야 한
 # 마지막 theme는 범례로서 위치를 지정해준다.
```

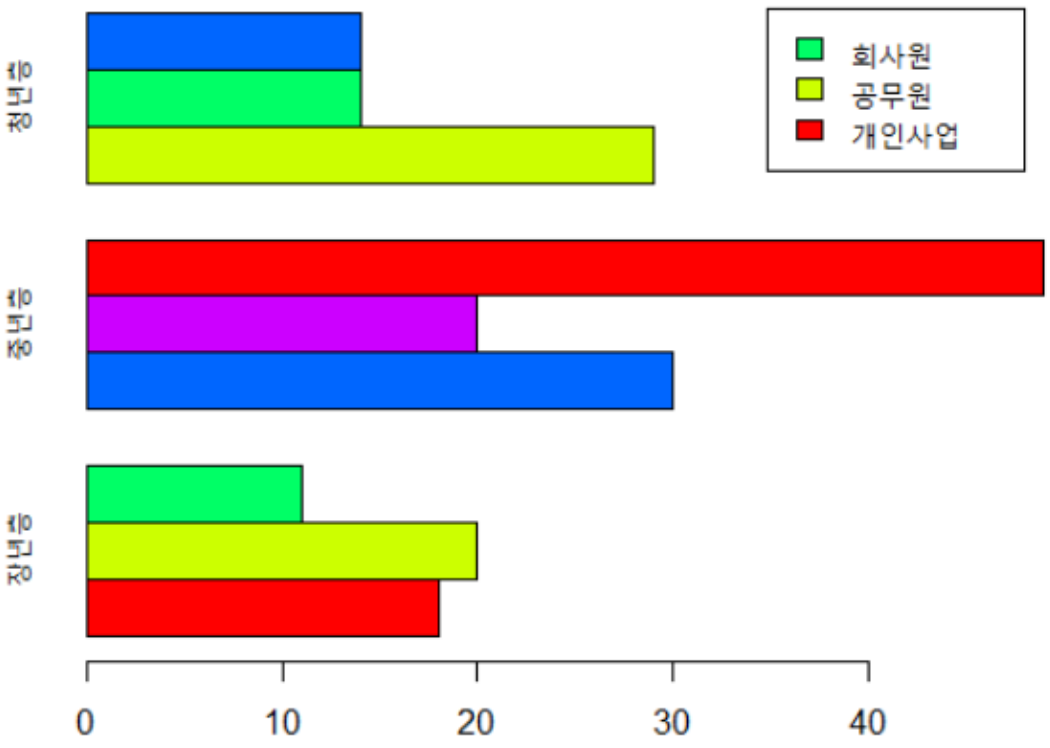
```r
#ggplot
#변환
resident_gender2←data.frame(resident2 = dataset$resident2,age2 = dataset$age2)
resident_gender_df ← as.data.frame(resident_gender)


ggplot(resident_gender_df , aes(x=Freq, y=Var2, fill=Var1 ) )+
  geom_bar(stat = "identity",position='dodge')+
  theme(legend.position = c(.8, .90))

ggplot(resident_gender_df , aes(x=Freq, y=Var2, fill=Var1 ) )+
  geom_col(position='dodge')
```



예제

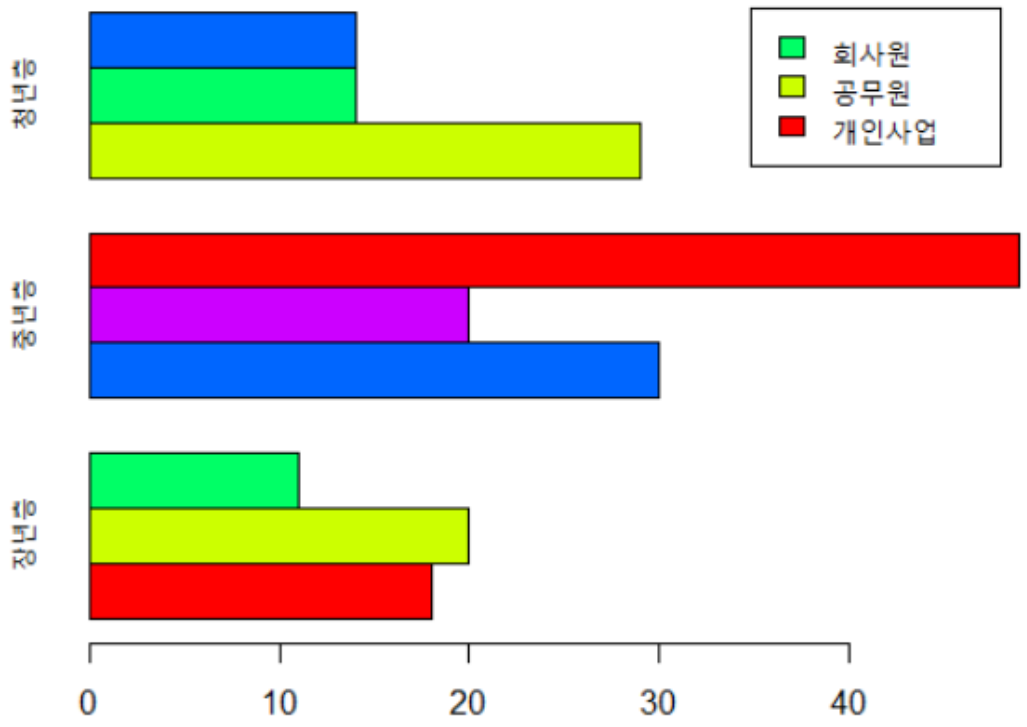
#barplot

```r
#예제
#직업유형(job2) vs 나이(age2)

class(dataset$job2)

dataset$job2 <- factor(dataset$job2)
class(dataset$age2)

job_age_table <- table(dataset$job2,dataset$age2)

#barplot
barplot(job_age_table,
        horiz = T,
        col = rainbow(5),
        beside = T,
        legend = row.names(job_age_table))

```



#ggplot

```r
#gplot

job_age_df ← as.data.frame(job_age_table)
class(job_age_df)
str(job_age_df)
names(job_age_df) ← c(Var1="직업", Var2="연령층", Freq = "나이")
str(job_age_df)


ggplot(job_age_df,aes(x=연령층,y=나이,fill=직업))+
  geom_col(position='dodge')+
  coord_flip()+
  theme(legend.position = c(.8, .90))
```



#실습 예제

```r
#실습예제

# 데이터 프레임의 복사본 생성하기
library(ggplot2)
midwest

midwest_raw<-as.data.frame(midwest)
midwest_new <- midwest_raw
str(midwest_new)
head(midwest_new)

# [문제]
# poptotal(전체인구) 변수를 total로,
# popasian(아시안 인구) 변수를 asian으로 수정하세요.
library(reshape)

midwest_new<-rename(midwest_new,
                    c(poptotal="total",
                      popasian="asian")
                    )

str(midwest_new)
```

```r
# [문제]
# total, asian 변수를 이용해 '전체 인구 대비 아시아 인구 백분율' percasian 파생변수를 만들고,
# 히스토그램을 만들어 도시들이 어떻게 분포하는지 살펴보세요.
head(midwest_new)
midwest_new$percasian2 <-midwest_new$asian/midwest_new$total
midwest_new$percasian2

midwest_new$county <- factor(midwest_new$county)

ggplot(midwest_new,aes(x=county,y=percasian2))+
  geom_col()

histogram(midwest_new$county)


# [문제]
# 아시아 인구 백분율 전체 평균을 구하고,
# 평균을 초과하면 "large",
# 그 외에는 "small"을 부여하는 mean 파생변수를 만들어 보세요.
asianMean<-mean(midwest_new$percasian2)

for (i in 1:nrow(dataset_new)){
  if (midwest_new$percasian2[i] > asianMean ){
    midwest_new$mean[i] <- "large"
  }
  else{
    midwest_new$mean[i] <- "small"
  }
}
midwest_new$mean

```

```r
# [문제]
# "large"와 "small"에 해당하는 지역이 얼마나 되는지 빈도표와
# 빈도 막대 그래프를 만들어 확인해 보세요.

ggplot(midwest_new , aes(x=mean,fill=mean))+
  geom_bar()


# ggplot2의 midwest 데이터를 사용하여 데이터 분석을 실습하는 문제 입니다.

# popadults는 해당 지역의 성인 인구,
# poptotal은 전체 인구를 나타냅니다.

# 1번 문제
# midwest 데이터에 '전체 인구 대비 미성년 인구 백분율' 변수를 추가하세요.
midwest_new$perYoungAge <-( midwest_new$total- midwest_new$popadults) / midwest_new$total
midwest_new$perYoungAge
class(midwest_new$perYoungAge)

# 2번 문제
# 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의
# 미성년 인구 백분율을 출력하시오.
sortPerYoung <- sort(midwest_new$perYoungAge, decreasing = TRUE)
sortPerYoung
?tail
tail(sortPerYoung,addrownums=4)
youngRange <- range(sortPerYoung[1:5])
youngRange[2]

value<-subset(midwest_new , midwest_new$perYoungAge >=youngRange[1] & midwest_new$perYoungAge <=youngRange[2], select=c(county,perYoungAge))
arrange(value,desc(value$perYoungAge))
```

```r
# 3번 문제
# 다음과 같은 분류표의 기준에 따라 미성년 비율 등급 변수를 추가하고,
# 각 등급에 몇 개의 지역이 있는지 알아보세요.

# 분류     기준
# large    40%이상
# middle   30 ~ 40미만
# small    30미만

for (i in 1:nrow(dataset_new)){
  if (midwest_new$perYoungAge[i] >= 0.4 ){
    midwest_new$youngGrade[i] <- "large"
  }
  else if(midwest_new$perYoungAge[i] >= 0.3 ){
    midwest_new$youngGrade[i] <- "middle"
  }
  else{⌨}
}
midwest_new$youngGrade

# 4번 문제
# popasian은 해당 지역의 아시아인 인구를 나타냅니다.
# '전체 인구 대비 아시아인 인구 백분율' 변수를 추가하고
# 하위 10개 지역의 state(주), county(지역), 아시아인 인구 백분율을 출력하세요.

str(midwest_new$asian)
midwest_new$asianVsTotal <- (midwest_new$asian / midwest_new$total)
midwest_new$asianVsTotal
asianRange=(sort(midwest_new$asianVsTotal))[1:10]
asianRange
library(dplyr)
dplyr::select(midwest_new,state,county,(sort(midwest_new$asianVsTotal))[1:10])
?dplyr::select
```

```r

value2<-subset(midwest_new, asianVsTotal >= range(asianRange)[1] & asianVsTotal <= range(asianRange)[2]  ,select=c(state,county,asianVsTotal) )


arrange(value2,value2$asianVsTotal)

midwest_new %>%
  subset(asianVsTotal >= range(asianRange)[1] & asianVsTotal <= range(asianRange)[2]) %>%
  select(c(state,county,asianVsTotal,county))
```

'R' 카테고리의 다른 글□

[R] 예제를 통한 데이터 전처리 작업□

[R] R을 통해 시계열 그래프 만들기 (자료 분석을 위한 시각화와 실습 예제)□

[R] 같은 형태의 ggplot 과 barplot 만들기 (차이 비교)□

[R] ggplot2 패키지 설치 에러시 해결 방법□

[R] R 을 활용한 데이터 탐색(Exploratory Data Analysis)□

[R] R ggplot 사용법 (데이터 시각화 도구)□

barplot barplot 함수 ggplot ggplot 함수

꾸까꾸

혼자 끄적끄적하는 블로그 입니다.