

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

노트북: [TIL-MY]

만든 날짜: 2020-08-01 오전 12:08

URL: <https://continuous-development.tistory.com/46?category=793392>

나무늘보의 개발 블로그

홈

태그

```
>
> 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 ...
R $ 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 51 45 62 57 36 37 29 35 56 ...
 $ position : int 4 3 1 1 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 ...
 $ 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/ 9 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.시군군"
```

ggplot과 barplot이 헷갈리는 경우가 있어 한번 정리했다.

분류 전체보기 

Python

Database

ASP.NET

Algorithm 

Deep learning

```
> #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$resident2)
[1] "1.서울특별시" "2.인천광역시" "3.대전광역시" "4.대구광역시" "5.서구군"
> levels(dataset$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.서구군         7    19     8
> class(resident_gender)
[1] "table"
> |
```

barplot

데이터는 벡터 또는 행렬로 받으면 된다.

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

AWS

ETC..

R 

공지사항

글 보실 때 주의사
항

: 최근글 : 인
기글

[Algorithm] 파이썬
을 파..

2020.07.31

[R] R
을 ...



2020.07.30

[R]
같...

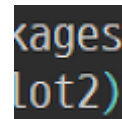


2020.07.30

[Algorithm] 프로그
래머스..

2020.07.30

[R]
gg...



2020.07.30

최근댓글

태그

ggplot 함수,

행렬, substr,

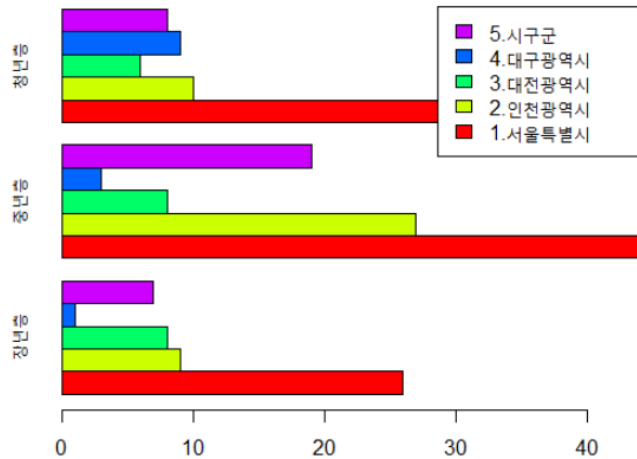
프로그래머스 스

킬체크,

```

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. 거
기에 x축을 빈도 y축을 연령층으로 잡고 색깔
구분을 지역으로 한다.
ggplot(resident_gender_df , aes(x=Freq, y=
Var2, fill=Var1 ))+
  geom_bar(stat = "identity",position='dod
ge')+ # 그 후 geom_bar 명령어를 써서 막대
그래프를 그리는데 geom_bar는
theme(legend.position = c(.8, .90))

```

프로그래머스 레
벨1,
barplot 함수,
AWS, 인스턴스,
ggplot2 패키지
설치 에러,
파이썬을 파이썬
답게,
cbind,
프로그래머스 스
킬체크 레벨1,
Oracle,
프로그래머스 레
벨1 python,
Python,
시계열 그래프,
시계열 그래프 그
리기,
ggplot2 에러,
DDL,
R 시계열 그래프,
ggplot, rbind,
테이블 생성,
날짜함수,
R 시계열 그래프
만들기,
SQL, 설정,
사용법, 파이썬,
Oracle SQL

전체 방문자

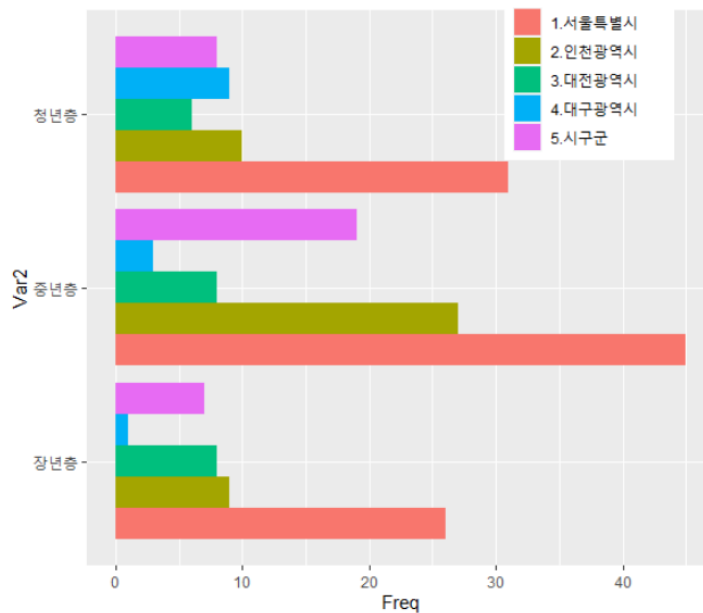
125

Today : 0

Yesterday : 2

기존에 하나의 축만가능해서 stat='identity'를 사용해야 한다. 그리고 position 을 통해 beside 같은 효과를 낸다.
마지막 theme는 범례로서 위치를 지정해 준다.

```
39 #ggplot
40 #변환
41 resident_gender2<-data.frame(resident2 = dataset$resident2,age2 = dataset$age2)
42 resident_gender_df <- as.data.frame(resident_gender)
43
44
45 ggplot(resident_gender_df , aes(x=Freq, y=Var2, fill=Var1 ) )+
46   geom_bar(stat = "identity",position='dodge')+
47   theme(legend.position = c(.8, .90))
48
49 ggplot(resident_gender_df , aes(x=Freq, y=Var2, fill=Var1 ) )+
50   geom_col(position='dodge')
51
```



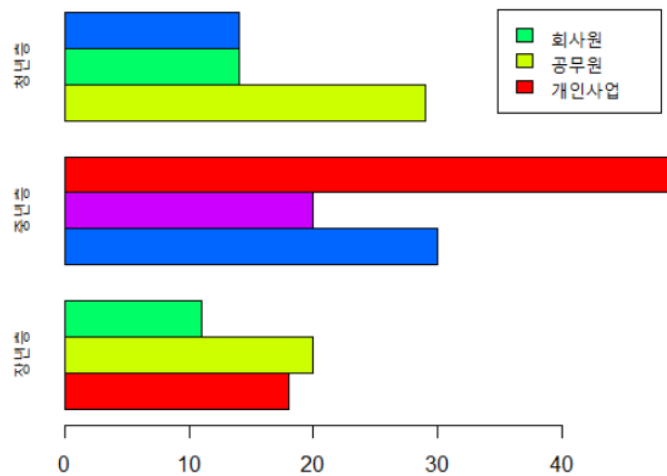
예제

#barplot

```

55 #예제
56 #직업유형(job2) vs 나이(age2)
57
58 class(dataset$job2)
59
60 dataset$job2 <- factor(dataset$job2)
61 class(dataset$age2)
62
63 job_age_table <- table(dataset$job2,dataset$age2)
64
65 #barplot
66 barplot(job_age_table,
67         horiz = T,
68         col = rainbow(5),
69         beside = T,
70         legend = row.names(job_age_table))
71

```

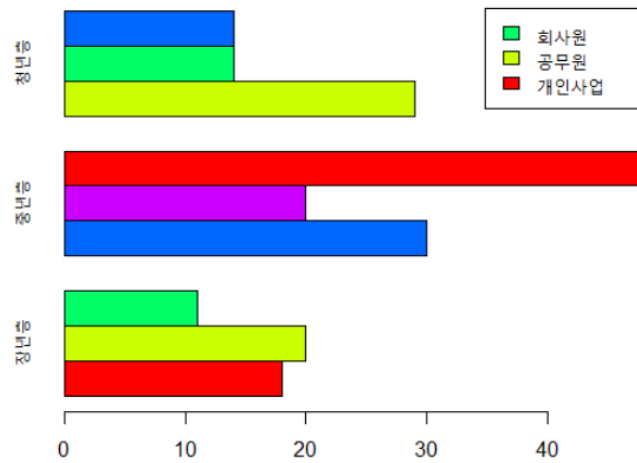


#ggplot

```

71
72 #ggplot
73
74 job_age_df <- as.data.frame(job_age_table)
75 class(job_age_df)
76 str(job_age_df)
77 names(job_age_df) <- c(Var1="직업", Var2="연령층", Freq = "나이")
78 str(job_age_df)
79
80
81 ggplot(job_age_df,aes(x=연령층,y=나이,fill=직업))+
82   geom_col(position='dodge')+
83   coord_flip()+
84   theme(legend.position = c(.8, .90))
85

```



#실습 예제

```

119
120 #실습예제
121
122 # 데이터 프레임의 복사본 생성하기
123 library(ggplot2)
124 midwest
125
126 midwest_raw<-as.data.frame(midwest)
127 midwest_new <- midwest_raw
128 str(midwest_new)
129 head(midwest_new)
130
131 # [[문제]]
132 # poptotal(전체인구) 변수를 total로,
133 # popasian(아시안 인구) 변수를 asian으로 수정하세요.
134 library(reshape)
135
136 midwest_new<-rename(midwest_new,
137                     c(poptotal="total",
138                       popasian="asian")
139                     )
140
141 str(midwest_new)

```

```

144 # [문제]
145 # total, asian 변수를 이용해 '전체 인구 대비 아시아 인구 백분율' percasian 파생변수를 만들고,
146 # 히스토그램을 만들어 도시들이 어떻게 분포하는지 살펴보세요.
147 head(midwest_new)
148 midwest_new$percasian2 <- midwest_new$asian / midwest_new$total
149 midwest_new$percasian2
150 |
151 midwest_new$county <- factor(midwest_new$county)
152
153 ggplot(midwest_new, aes(x=county, y=percasian2)) +
154   geom_col()
155
156 histogram(midwest_new$county)
157
158 # [문제]
159 # 아시아 인구 백분율 전체 평균을 구하고,
160 # 평균을 초과하면 "large",
161 # 그 외에는 "small"을 부여하는 mean 파생변수를 만들어 보세요.
162 asianMean <- mean(midwest_new$percasian2)
163
164 for (i in 1:nrow(dataset_new)){
165   if (midwest_new$percasian2[i] > asianMean){
166     midwest_new$mean[i] <- "large"
167   }
168   else{
169     midwest_new$mean[i] <- "small"
170   }
171 }
172 }
173 midwest_new$mean
174

```

```

176 # [문제]
177 # "large"와 "small"에 해당하는 지역이 얼마나 되는지 빈도표와
178 # 빈도 막대 그래프를 만들어 확인해 보세요.
179
180 ggplot(midwest_new, aes(x=mean, fill=mean)) +
181   geom_bar()
182
183 # ggplot2의 midwest 데이터를 사용하여 데이터 분석을 실습하는 문제입니다.
184
185 # popadults는 해당 지역의 성인 인구,
186 # poptotal은 전체 인구를 나타냅니다.
187
188 # 1번 문제
189 # midwest 데이터에 '전체 인구 대비 미성년 인구 백분율' 변수를 추가하세요.
190 midwest_new$perYoungAge <- (midwest_new$total - midwest_new$popadults) / midwest_new$total
191 midwest_new$perYoungAge
192 class(midwest_new$perYoungAge)
193
194 # 2번 문제
195 # 미성년 인구 백분율이 가장 높은 상위 5개 county(지역)의
196 # 미성년 인구 백분율을 출력하세요.
197 sortPerYoung <- sort(midwest_new$perYoungAge, decreasing = TRUE)
198 sortPerYoung
199 ?tail
200 tail(sortPerYoung, addn=nrow-4)
201 youngRange <- range(sortPerYoung[1:5])
202 youngRange[2]
203
204 value <- subset(midwest_new, midwest_new$perYoungAge > youngRange[1] & midwest_new$perYoungAge < youngRange[2], select=c(county, perYoungAge))
205 arrange(value, desc(value$perYoungAge))
206

```

```

208 # 3번 문제
209 # 다음과 같은 분류표의 기준에 따라 미성년 비율 등급 변수를 추가하고,
210 # 각 등급에 몇 개의 지역이 있는지 알아보세요.
211
212 # 분류 기준
213 # large 40%이상
214 # middle 30 ~ 40미만
215 # small 30미만
216
217 for (i in 1:nrow(dataset_new)){
218   if (midwest_new$perYoungAge[i] ≥ 0.4){
219     midwest_new$youngGrade[i] <- "large"
220   }
221   else if (midwest_new$perYoungAge[i] ≥ 0.3){
222     midwest_new$youngGrade[i] <- "middle"
223   }
224   else{ }
225 }
226 midwest_new$youngGrade
227
228 # 4번 문제
229 # popasian은 해당 지역의 아시아인 인구를 나타냅니다.
230 # '전체 인구 대비 아시아인 인구 백분율' 변수를 추가하고
231 # 하위 10개 지역의 state(주), county(지역), 아시아인 인구 백분율을 출력하세요.
232
233 str(midwest_new$asian)
234 midwest_new$asianVsTotal <- (midwest_new$asian / midwest_new$total)
235 midwest_new$asianVsTotal
236 asianRange <- (sort(midwest_new$asianVsTotal))[1:10]
237 asianRange
238 library(dplyr)
239 dplyr::select(midwest_new, state, county, (sort(midwest_new$asianVsTotal))[1:10])
240 ?dplyr::select
241

```

```

243 value2 <- subset(midwest_new, asianVsTotal ≥ range(asianRange)[1] & asianVsTotal ≤ range(asianRange)[2], select=c(state, county, asianVsTotal))
244
245 arrange(value2, value2$asianVsTotal)
246
247 midwest_new %>%
248   subset(asianVsTotal ≥ range(asianRange)[1] & asianVsTotal ≤ range(asianRange)[2]) %>%
249   select(c(state, county, asianVsTotal, county))
250

```



'R' 카테고리의 다른 글

- [R] R을 통해 시계열 그래프 만들기
(자료 분석을 위한 시각화와 실습 예제) (0) 2020.07.30
- [R] 같은 형태의 ggplot 과 barplot
만들기 (차이 비교) (0) 2020.07.30
- [R] ggplot2 패키지 설치 에러시 해결
방법 (0) 2020.07.30
- [R] R 을 활용한 데이터 탐색(Exploratory Data Analysis) (0) 2020.07.30
- [R] R ggplot 사용법 (데이터 시각화
도구) (0) 2020.07.28
- [R] R 에서 사용되는 기본적인 시각
화 그래프-2 (0) 2020.07.28

태그

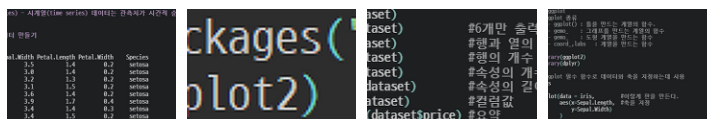
barplot

barplot 함수

ggplot

ggplot 함수

관련글



[R] R을 통... [R] ggplot... [R] R 을 활... [R] R ggpl...

댓글 0

TEL. 02.1234.5678 / 경기 성남시 분당구 판교역로
© Kakao Corp.

