

Bài giảng 8b: Phân tích bằng biểu đồ thanh (barplot)

Nguyễn Văn Tuấn

Garvan Institute of Medical Research, Australia
Ton Duc Thang University, Vietnam

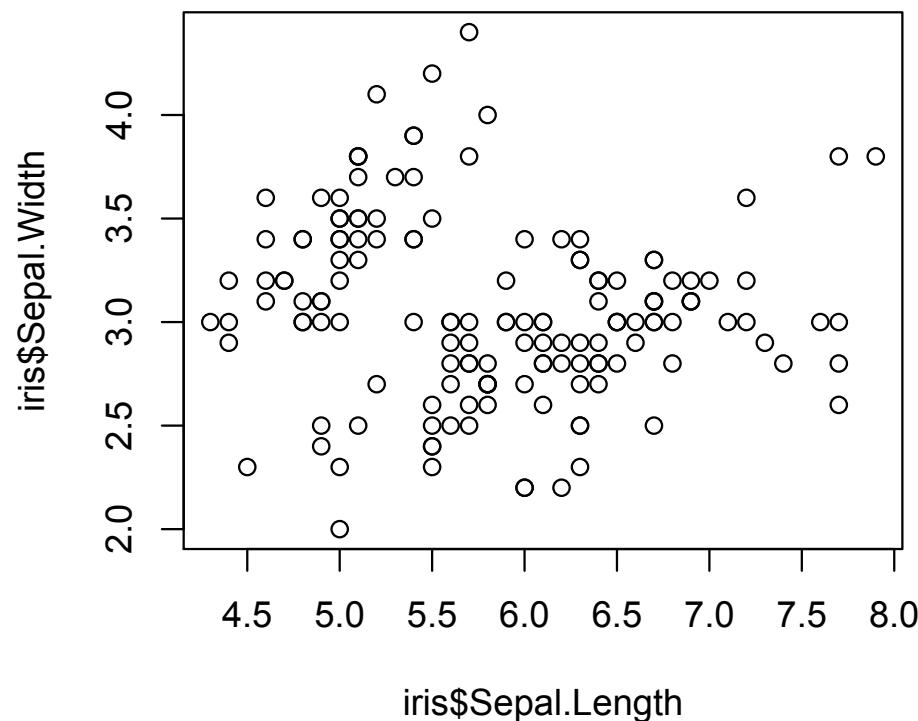
Nội dung

- Biểu đồ từ dữ liệu summary
- Biểu đồ từ dữ liệu thô (raw data)

**Dữ liệu tóm tắt (summary
data)**

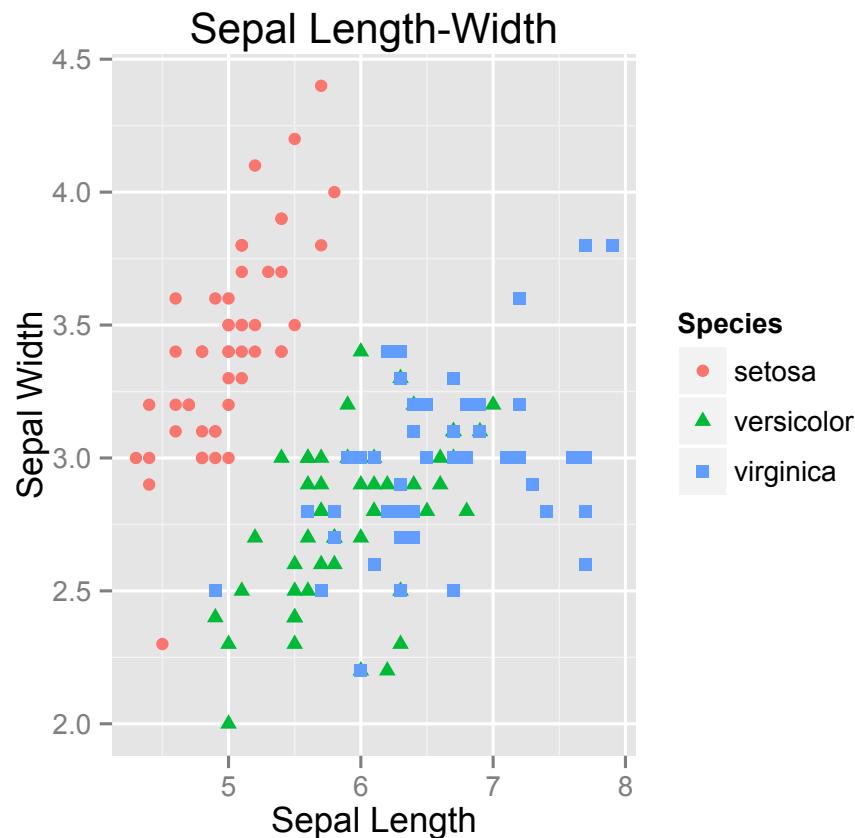
Một biểu đồ buồn chán ...

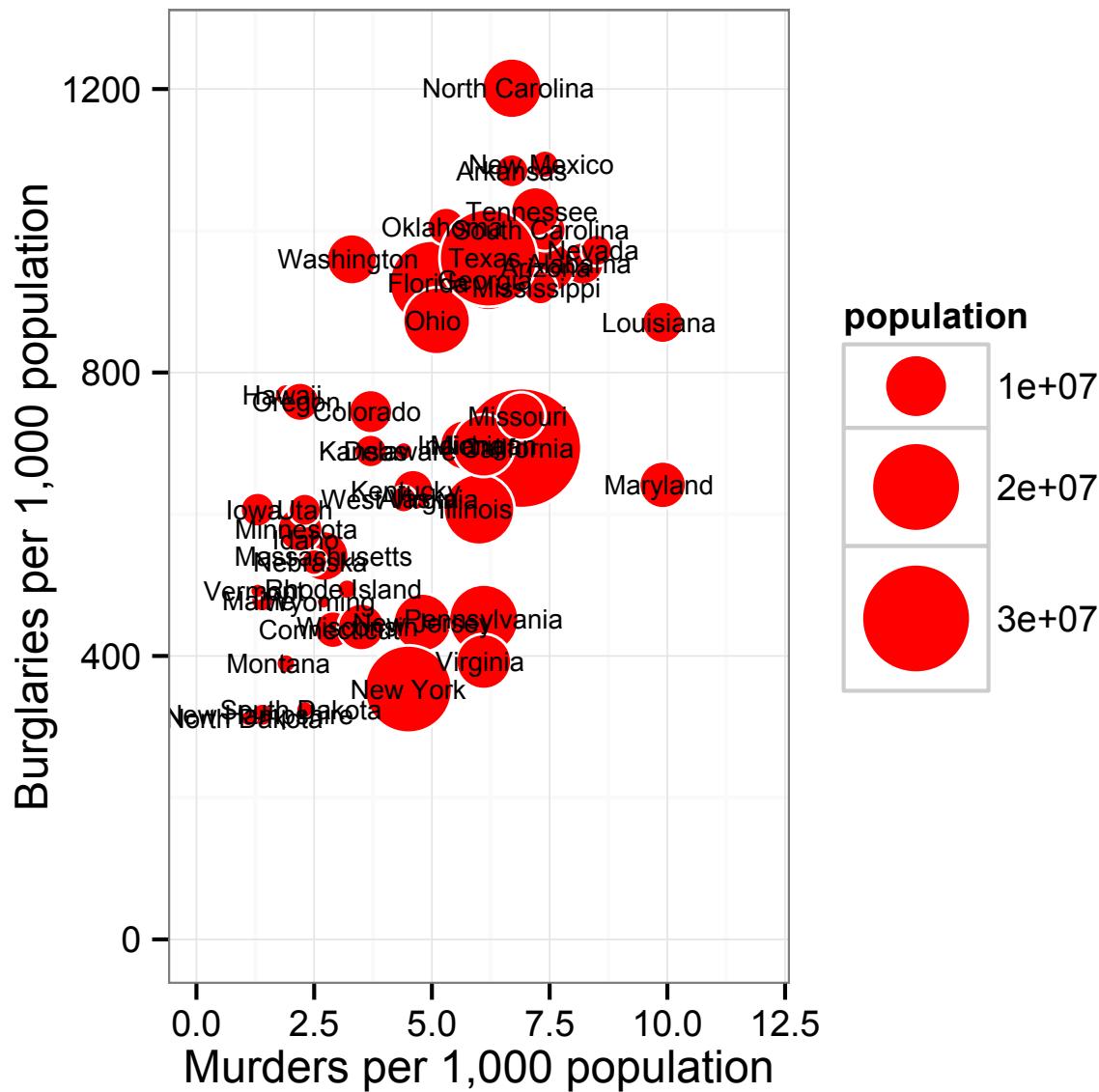
```
library(ggplot2);  
plot(x=iris$Sepal.Length, y=iris$Sepal.Width)
```



Sinh động hơn ...

```
p = ggplot(data=iris, aes(x = Sepal.Length, y = Sepal.Width))  
p + geom_point(aes(color=Species, shape=Species)) + xlab("Sepal Length") + ylab("Sepal Width") + ggttitle("Sepal Length-Width")
```





Ví dụ minh họa: hoa hậu Việt Nam

Name	YoB	Height	Weight
Bùi Bích Phượng	1971	1988	157
Nguyễn Diệu Hoa	1969	1990	158
Hà Kiều Anh	1976	1992	174
Nguyễn Thu Thủy	1976	1994	172
Nguyễn Thiên Nga	1976	1996	170
Nguyễn Thị Ngọc Khanh	1976	1998	172
Phan Thu Ngân	1980	2000	169
Phạm Thị Mai Phương	1985	2002	169
Nguyễn Thị Huyền	1985	2004	172
Mai Phương Thúy	1988	2006	181
Trần Thị Thùy Dung	1990	2008	182
Đặng Thị Ngọc Hân	1989	2010	173
Đặng Thu Thảo	1991	2012	173
Nguyễn Cao Kỳ Duyên	1996	2014	173

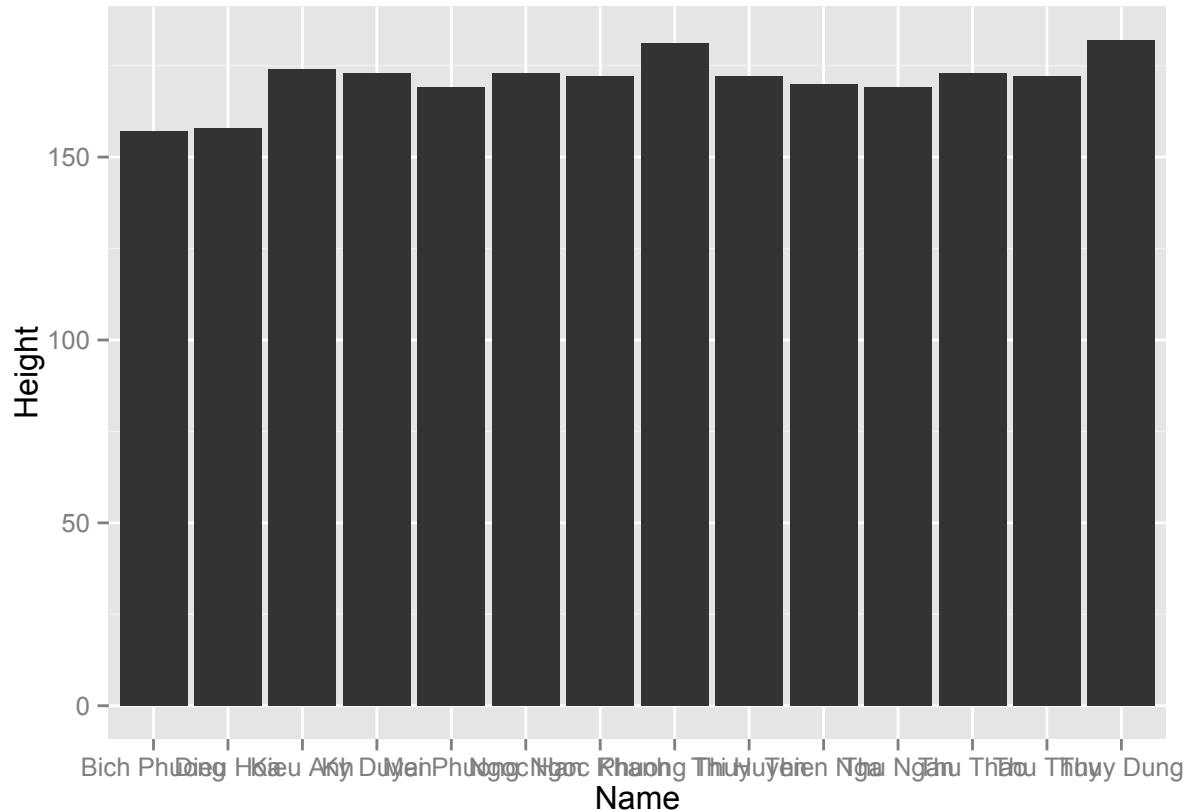
Dữ liệu về tỉ lệ nghèo một số tỉnh

```
Name = c("Bich Phuong", "Dieu Hoa", "Kieu Anh", "Thu  
Thuy", "Thien Nga", "Ngoc Khanh", "Thu Ngan", "Mai  
Phuong", "Thi Huyen", "Phuong Thuy", "Thuy Dung", "Ngoc  
Han", "Thu Thao", "Ky Duyen")  
  
YoB = c(1971, 1969, 1976, 1976, 1976, 1976, 1980, 1985,  
1985, 1988, 1990, 1989, 1991, 1996)  
  
Height = c(157, 158, 174, 172, 170, 172, 169, 169, 172,  
181, 182, 173, 173, 173)  
  
Weight = c(50, NA, NA, NA, NA, 50, 49, 49, 52, 60, 61.5,  
55, 49, 49)  
  
hoahau = data.frame(Name, YoB, Height, Weight)
```

Biểu đồ đơn giản và cơ bản

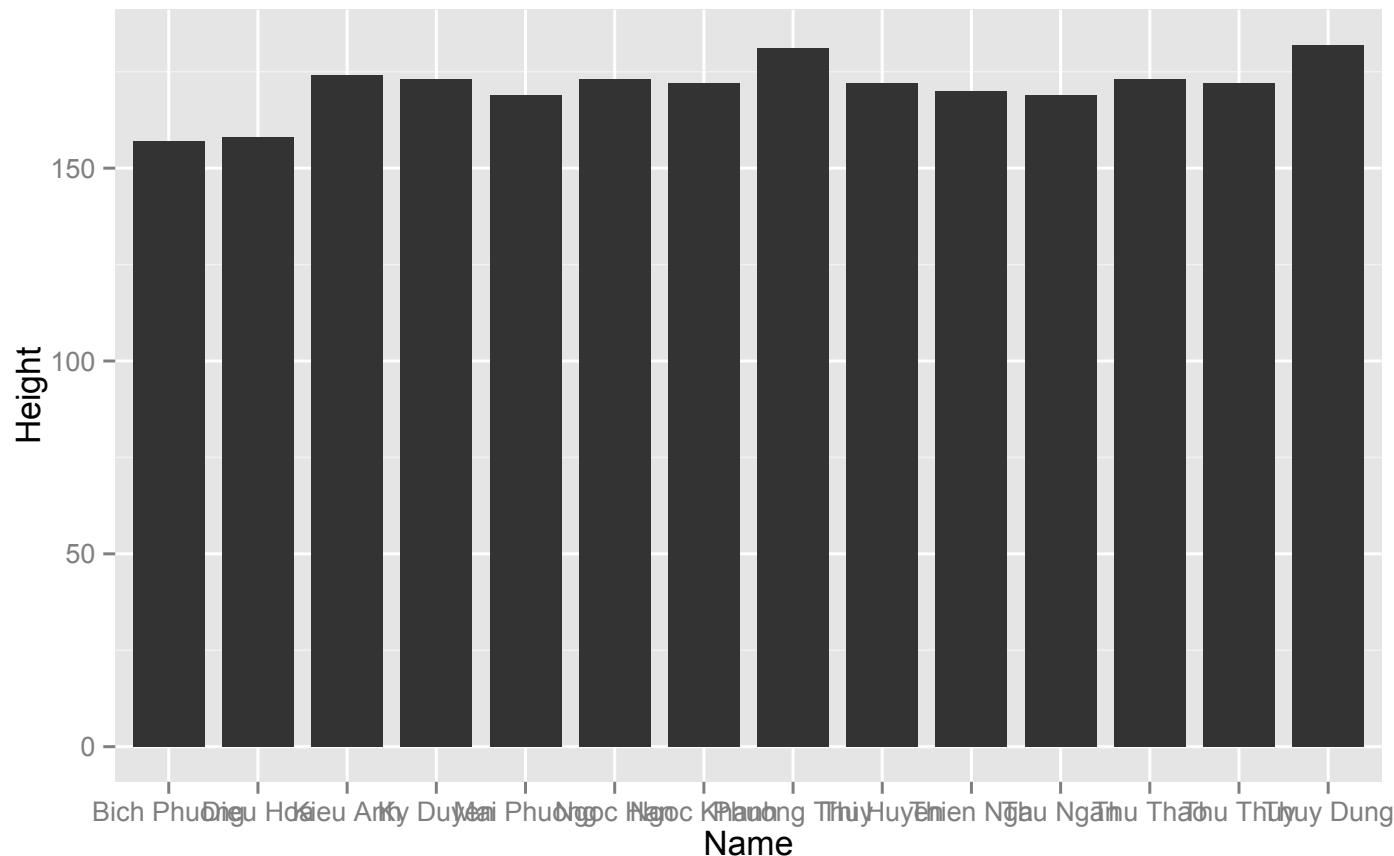
```
# basic plot
```

```
ggplot(data=hoahau, aes(x=Name, y=Height)) +  
geom_bar(stat="identity")
```



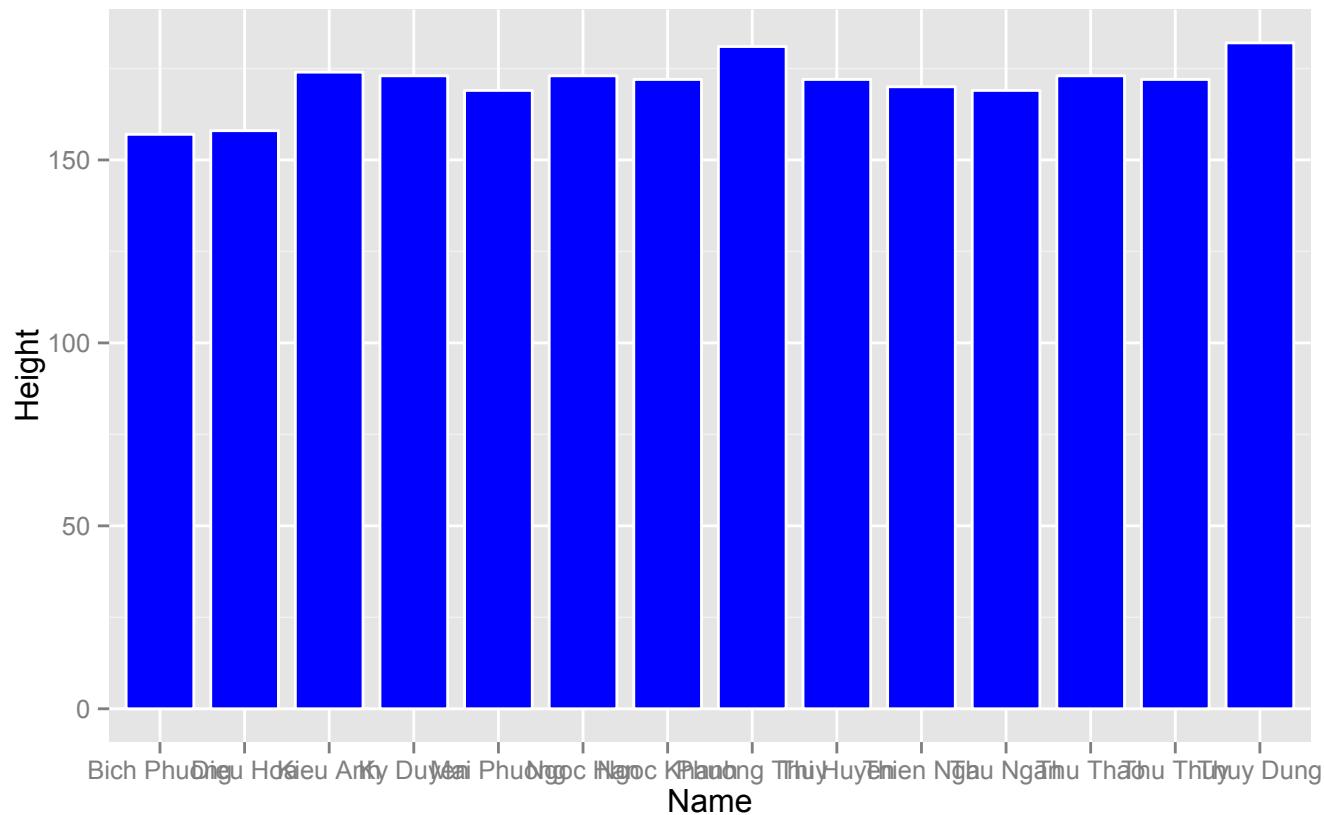
Biểu đồ đơn giản và cơ bản

```
ggplot(data=hoahau, aes(x=Name, y=Height)) +  
  geom_bar(stat="identity", width=0.8)
```



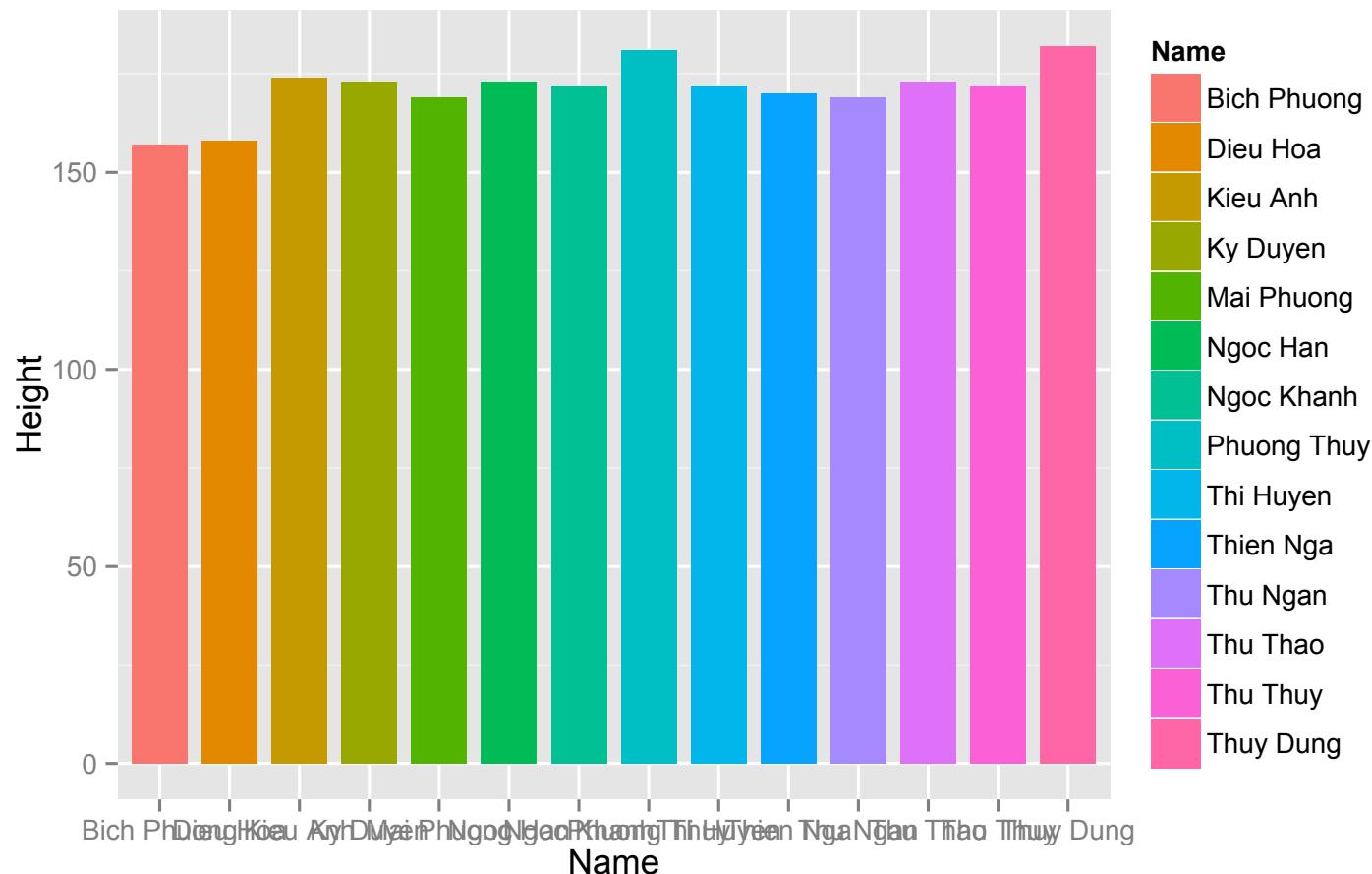
Thay đổi màu

```
ggplot(data=hoahau, aes(x=Name, y=Height)) +  
  geom_bar(stat="identity", width=0.8, colour="white",  
           fill="blue")
```



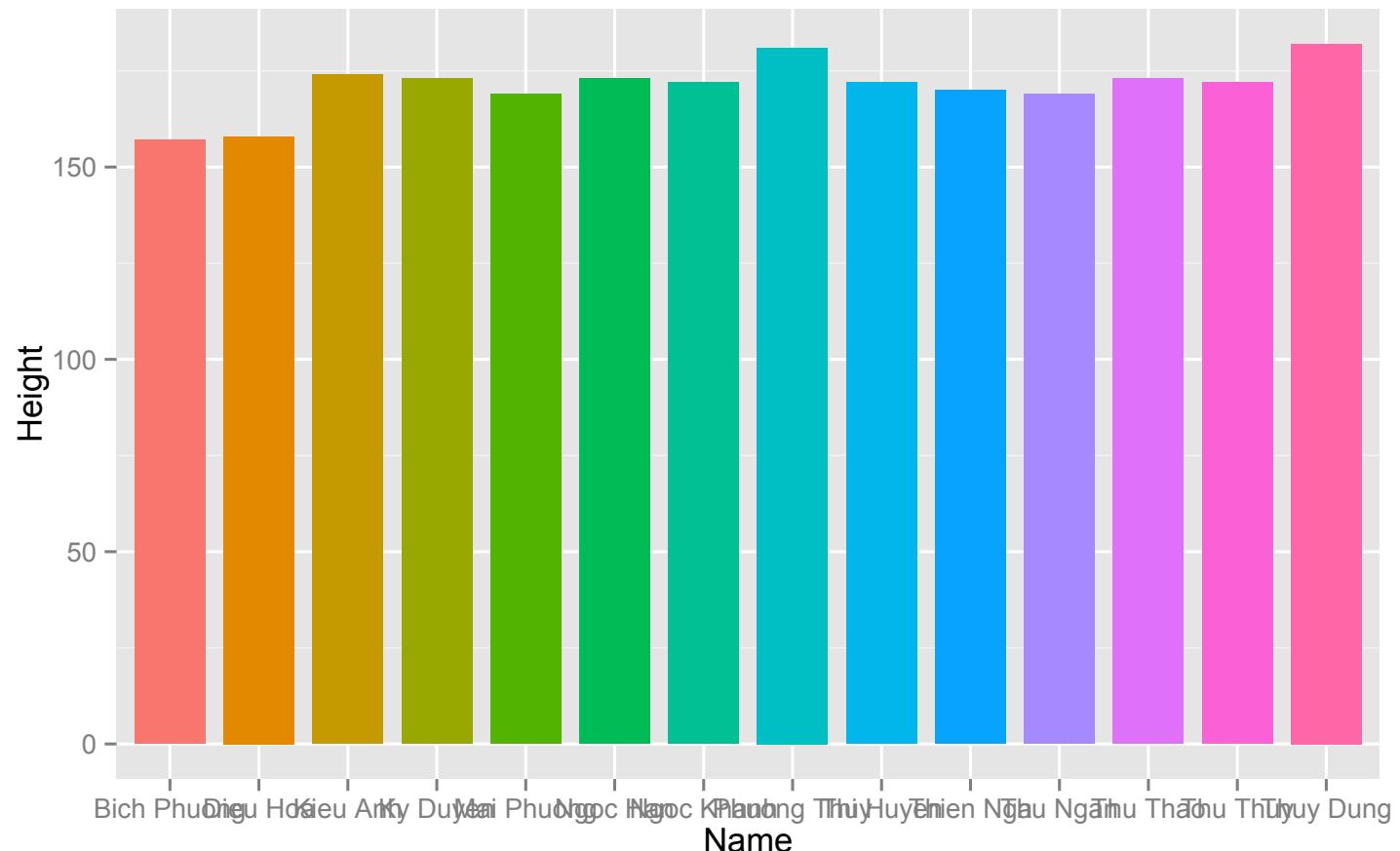
Thay đổi màu (nhiều màu)

```
ggplot(data=hoahau, aes(x=Name, y=Height, fill=Name)) +  
  geom_bar(stat="identity", width=0.8)
```

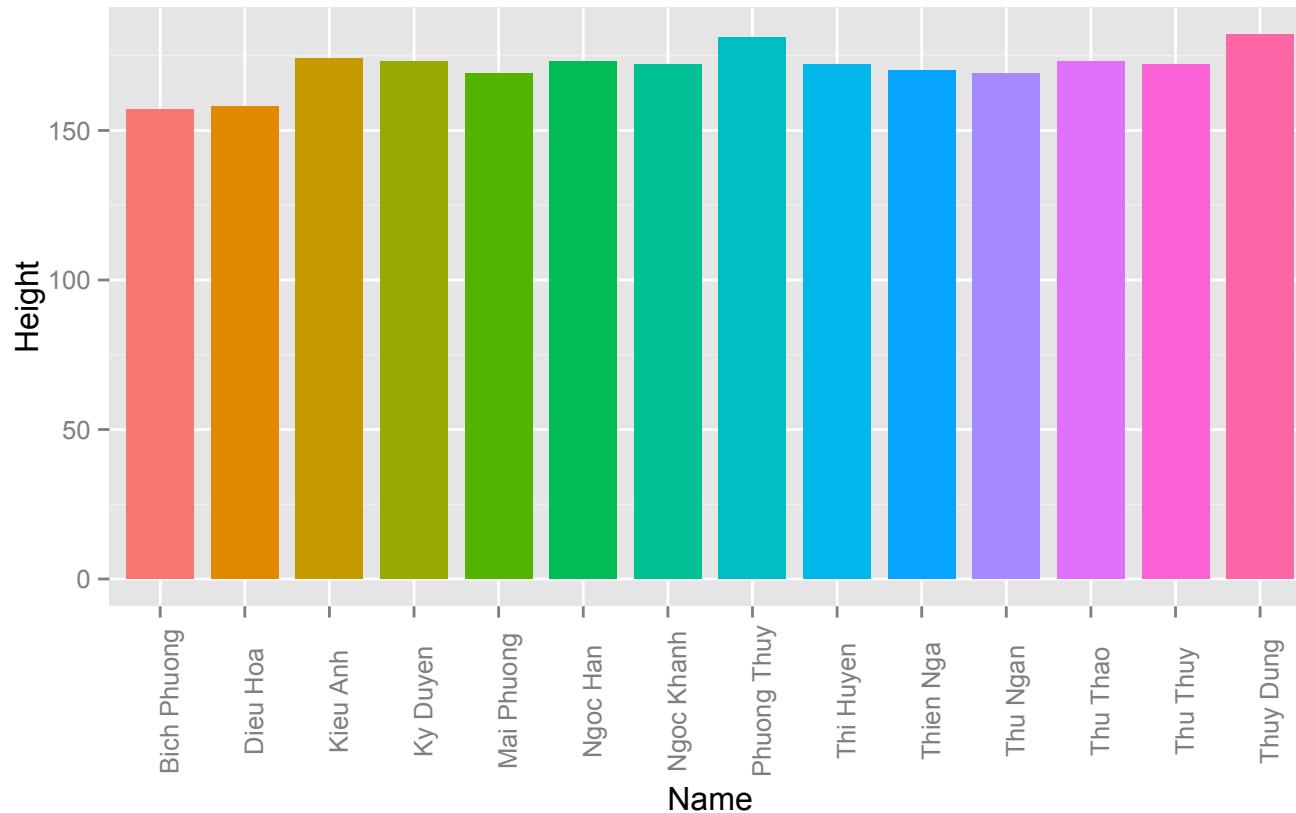


Vị trí legend

```
ggplot(data=hoahau, aes(x=Name, y=Height, fill=Name)) +  
  geom_bar(stat="identity", width=0.8) +  
  theme(legend.position="none")
```



```
p = ggplot(data=hoahau, aes(x=Name, y=Height,  
fill=Name))  
  
p = p + geom_bar(stat="identity", width=0.8)  
  
p + theme(legend.position="none",  
axis.text.x=element_text(angle=90))
```



```

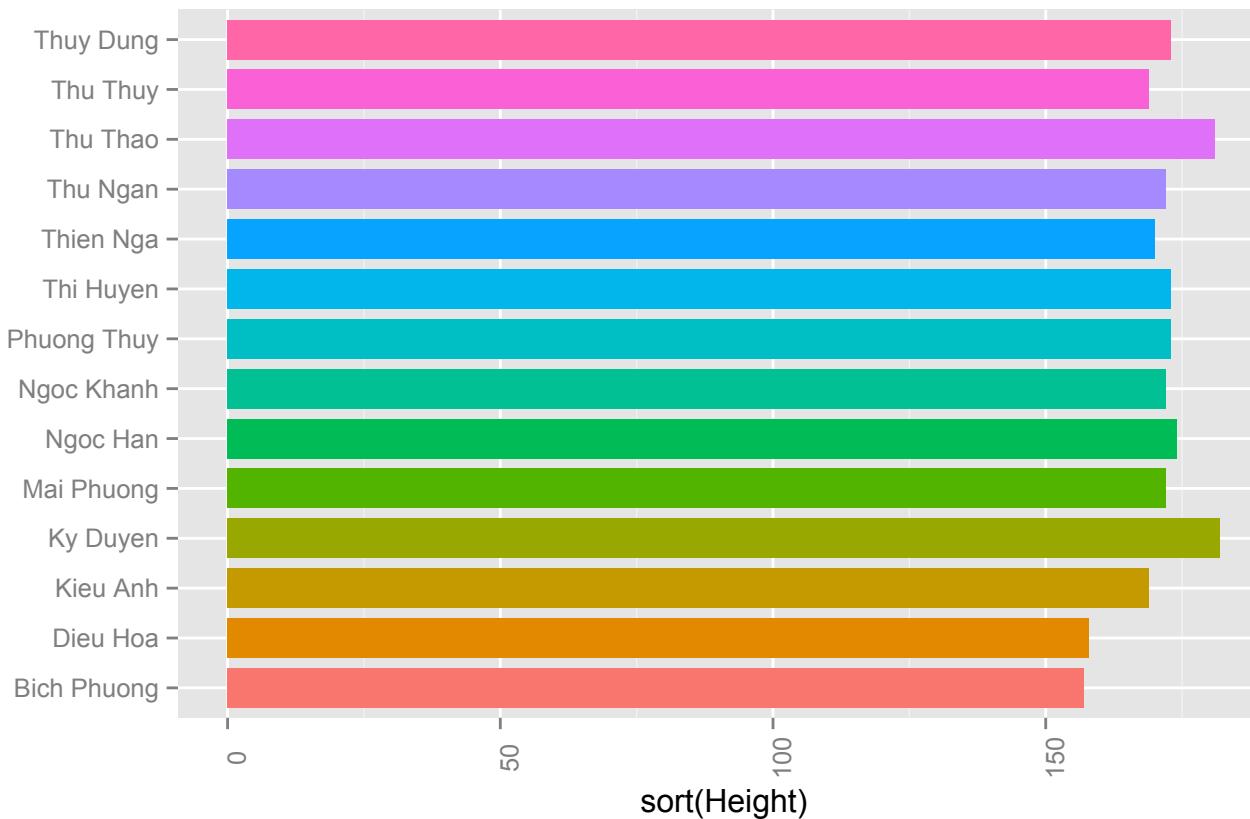
p = ggplot(data=hoahau, aes(x=Name, y=Height,
fill=Name))

p = p + geom_bar(stat="identity", width=0.8)

p = p + theme(legend.position="none",
axis.text.x=element_text(angle=90))

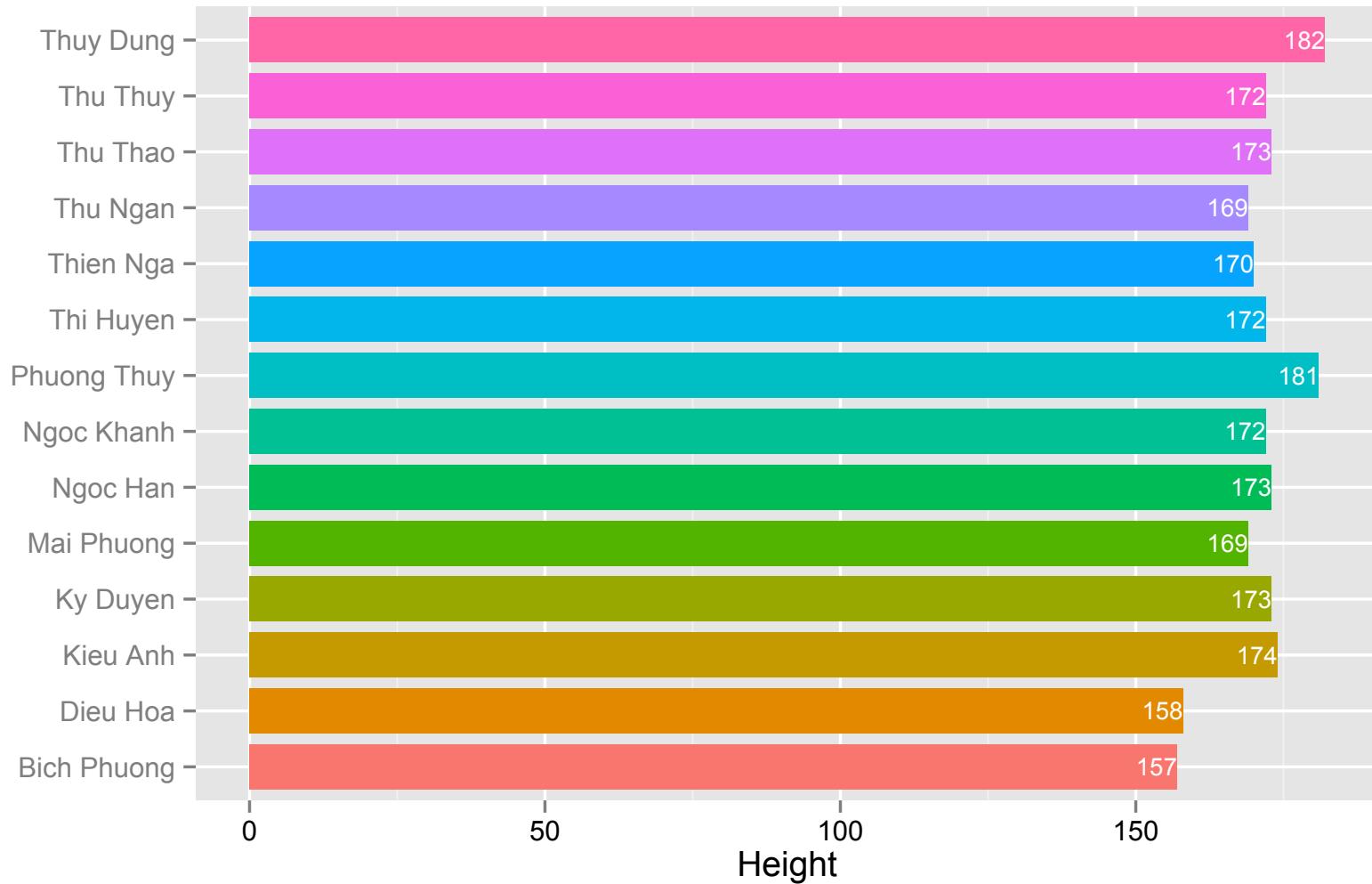
p + coord_flip() + xlab("")

```



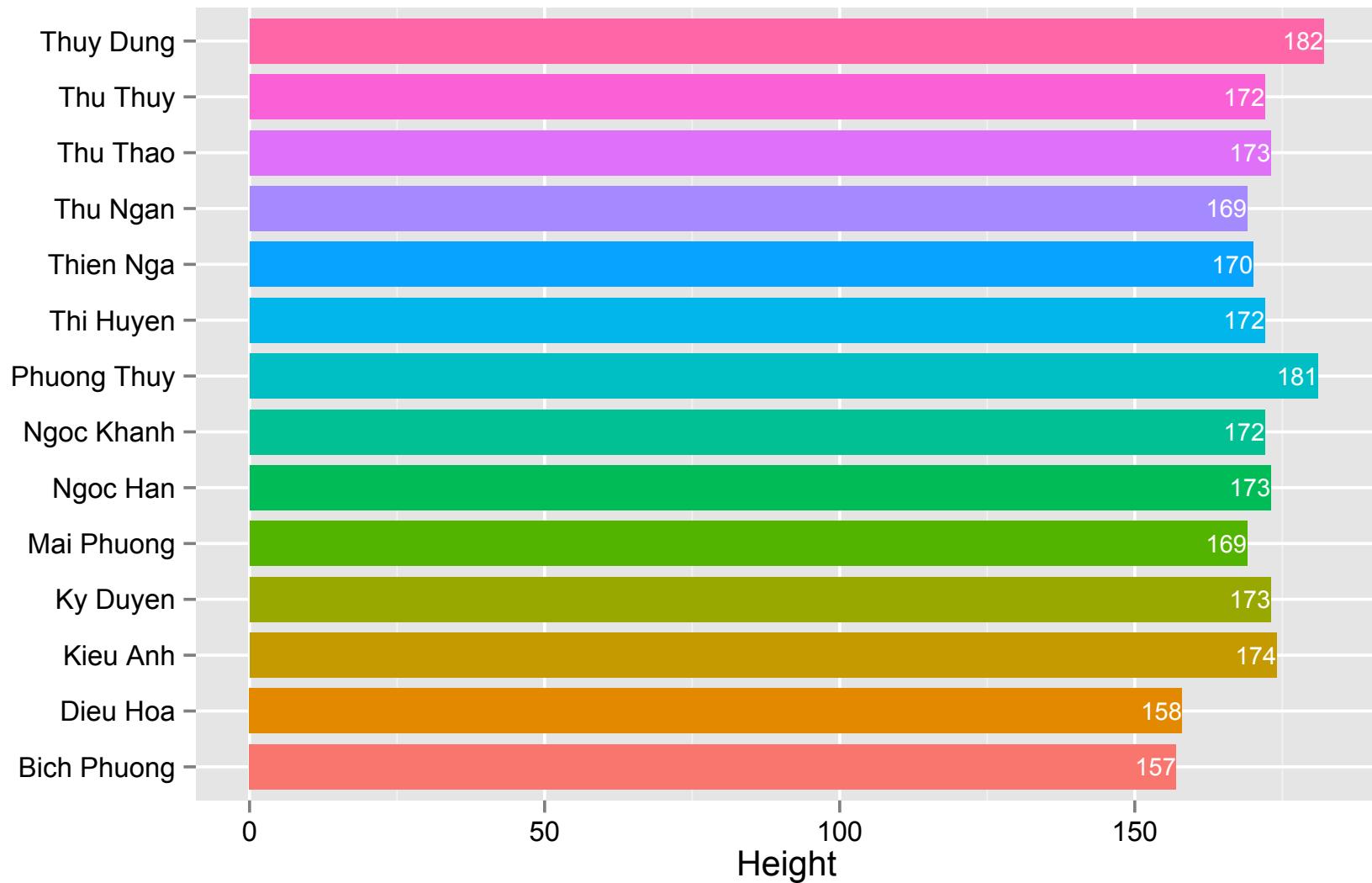
Thêm giá trị của mỗi bar

```
p = ggplot(data=hoahau, aes(x=Name, y=Height,  
fill=Name))  
  
p = p + geom_bar(stat="identity", width=0.8)  
  
p = p + theme(legend.position="none",  
axis.text.x=element_text(angle=0, color="black"))  
  
p = p + coord_flip() + xlab(" ")  
  
p + geom_text(aes(label=Height), hjust=1, size=3,  
color="white")
```



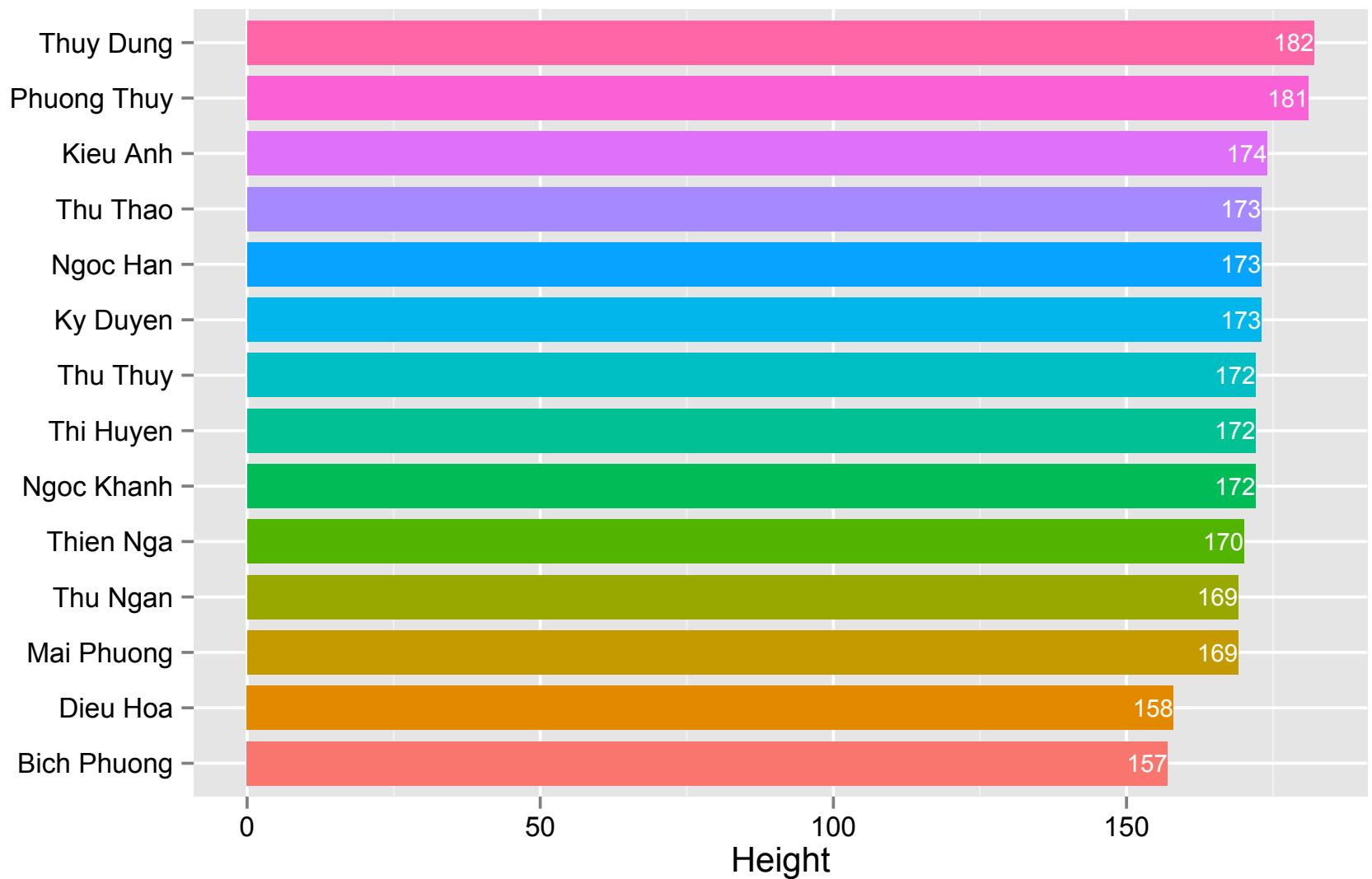
Thay màu trực tung

```
p = ggplot(data=hoahau, aes(x=Name, y=Height,  
fill=Name))  
  
p = p + geom_bar(stat="identity", width=0.8)  
  
p = p + theme(legend.position="none",  
axis.text.x=element_text(angle=0, color="black") ,  
axis.text.y=element_text(color="black"))  
  
p = p + coord_flip() + xlab(" ")  
  
p + geom_text(aes(label=Height), hjust=1, size=3,  
color="white")
```



Sắp xếp thứ tự theo giá trị

```
# Trước hết tạo ra một object khác gọi là hh, theo thứ  
tự chiều cao, dùng hàm transform  
  
hh = transform(hoahau, Name=reorder(Name, Height))  
  
# Vẽ  
  
p = ggplot(data=hh, aes(x=Name, y=Height, fill=Name))  
p = p + geom_bar(stat="identity", width=0.8)  
p = p + theme(legend.position="none",  
axis.text.x=element_text(angle=0, color="black"),  
axis.text.y=element_text(color="black"))  
p = p + coord_flip() + xlab(" ")  
p + geom_text(aes(label=Height), hjust=1, size=3,  
color="white")
```



Thay đổi theme

```
require(ggthemes)

p = ggplot(data=hh, aes(x=Name, y=Height, fill=Name))

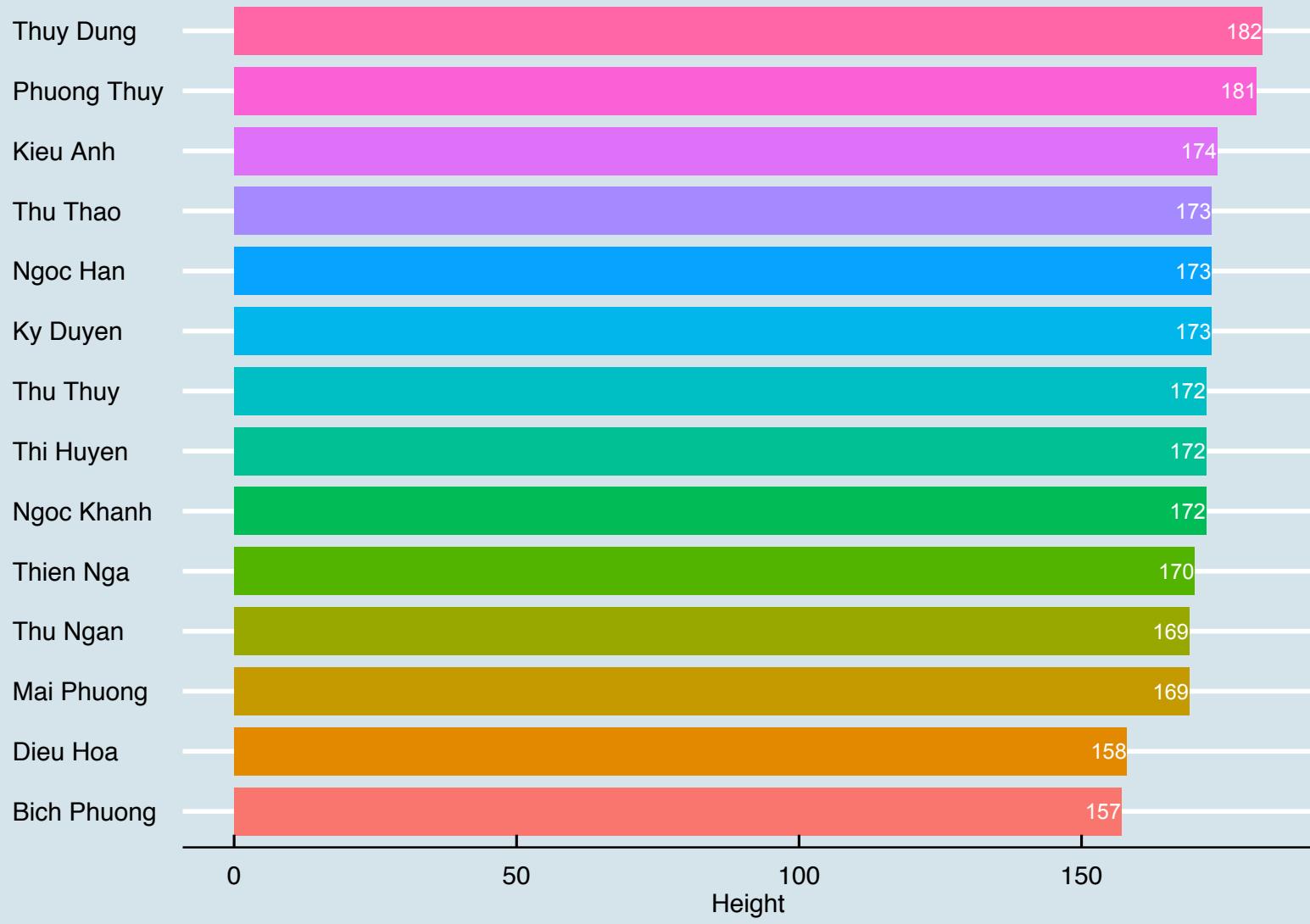
p = p + geom_bar(stat="identity", width=0.8)

p = p + theme(axis.text.x=element_text(angle=0,
color="black"), axis.text.y=element_text(size=15,
color="black"))

p = p + coord_flip() + xlab(" ")

p = p + geom_text(aes(label=Height), hjust=1, size=3,
color="white")

p + theme_economist() + theme(legend.position="none")
```



Dữ liệu gốc (raw data)

Dữ liệu PISA

```
setwd("~/Dropbox/World Bank 2014/Data for 2015  
workshop")  
  
pisa = read.csv("~/Dropbox/World Bank 2014/Data for 2015  
workshop/PISA DATA.csv", header=T)  
  
attach(pisa)  
  
# t=tabular(REGION ~ AREA*(n=1+Percent("col")))  
# html(t, "test.html")
```

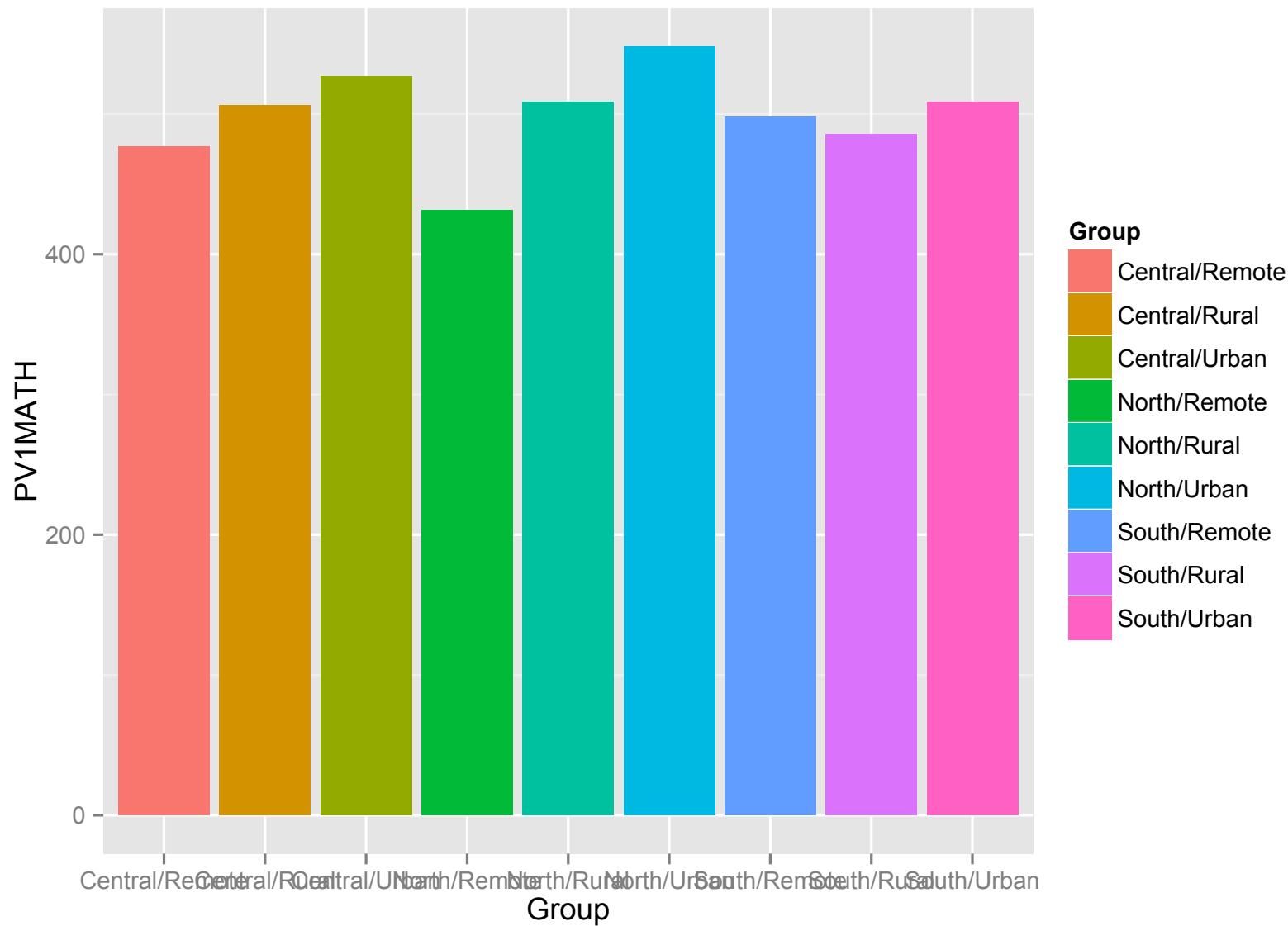
Mục tiêu

- Thể hiện điểm trung bình cho từng miền (REGION) và vùng (AREA)
- Cần thêm sai số chuẩn (standard error)
- Sắp xếp thứ tự theo điểm trung bình

Chuẩn bị data

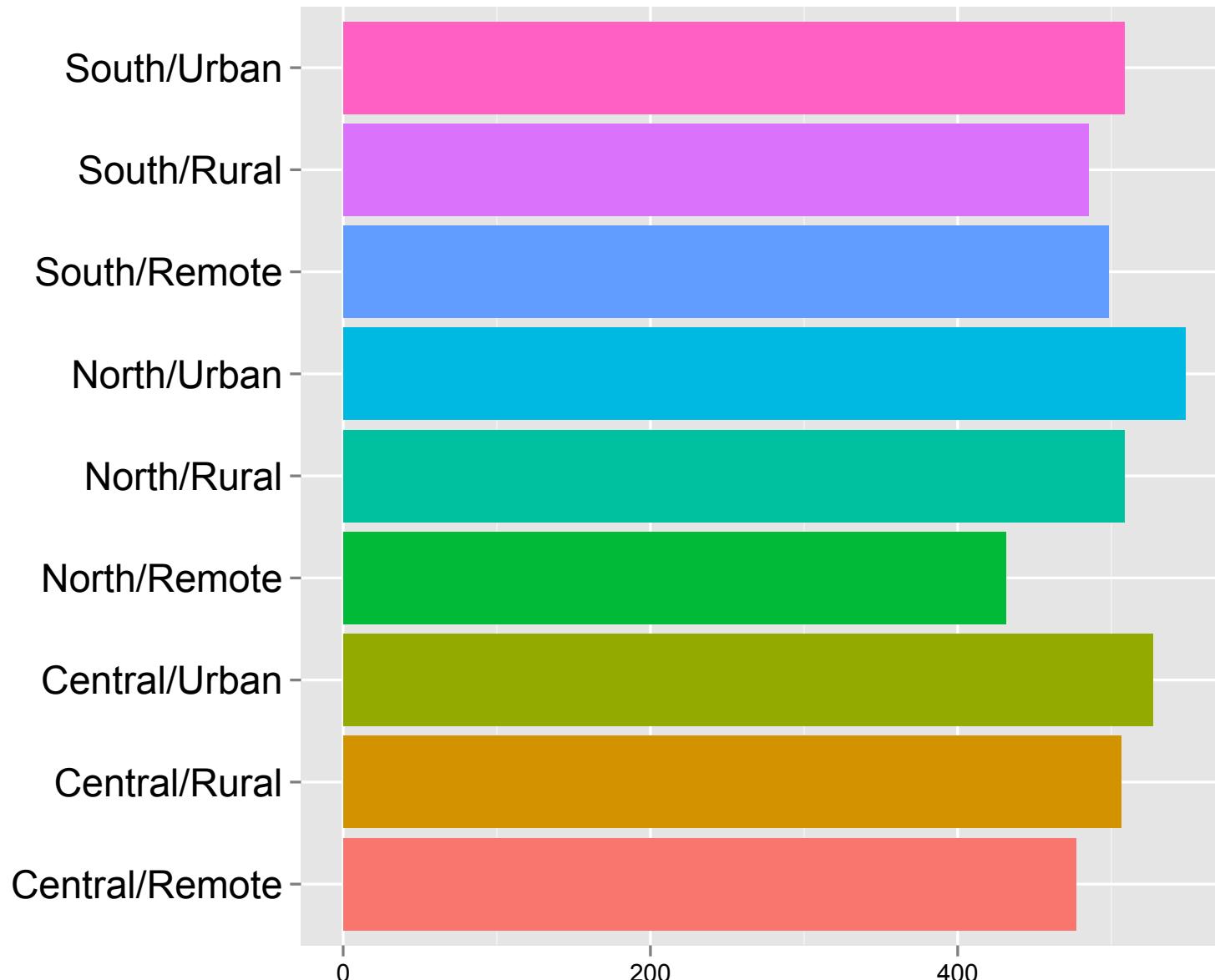
```
pisa$Group[pisa$REGION=="NORTH" & pisa$AREA=="URBAN"] = "North/Urban"
pisa$Group[pisa$REGION=="NORTH" & pisa$AREA=="RURAL"] = "North/Rural"
pisa$Group[pisa$REGION=="NORTH" & pisa$AREA=="REMOTE"] = "North/Remote"
pisa$Group[pisa$REGION=="CENTRAL" & pisa$AREA=="URBAN"] = "Central/Urban"
pisa$Group[pisa$REGION=="CENTRAL" & pisa$AREA=="RURAL"] = "Central/Rural"
pisa$Group[pisa$REGION=="CENTRAL" & pisa$AREA=="REMOTE"] = "Central/Remote"
pisa$Group[pisa$REGION=="SOUTH" & pisa$AREA=="URBAN"] = "South/Urban"
pisa$Group[pisa$REGION=="SOUTH" & pisa$AREA=="RURAL"] = "South/Rural"
pisa$Group[pisa$REGION=="SOUTH" & pisa$AREA=="REMOTE"] = "South/Remote"
```

```
attach(pisa)
p = ggplot(pisa, aes(x=Group, y=PV1MATH, fill=Group))
p = p + stat_summary(fun.y="mean", geom="bar")
p
```

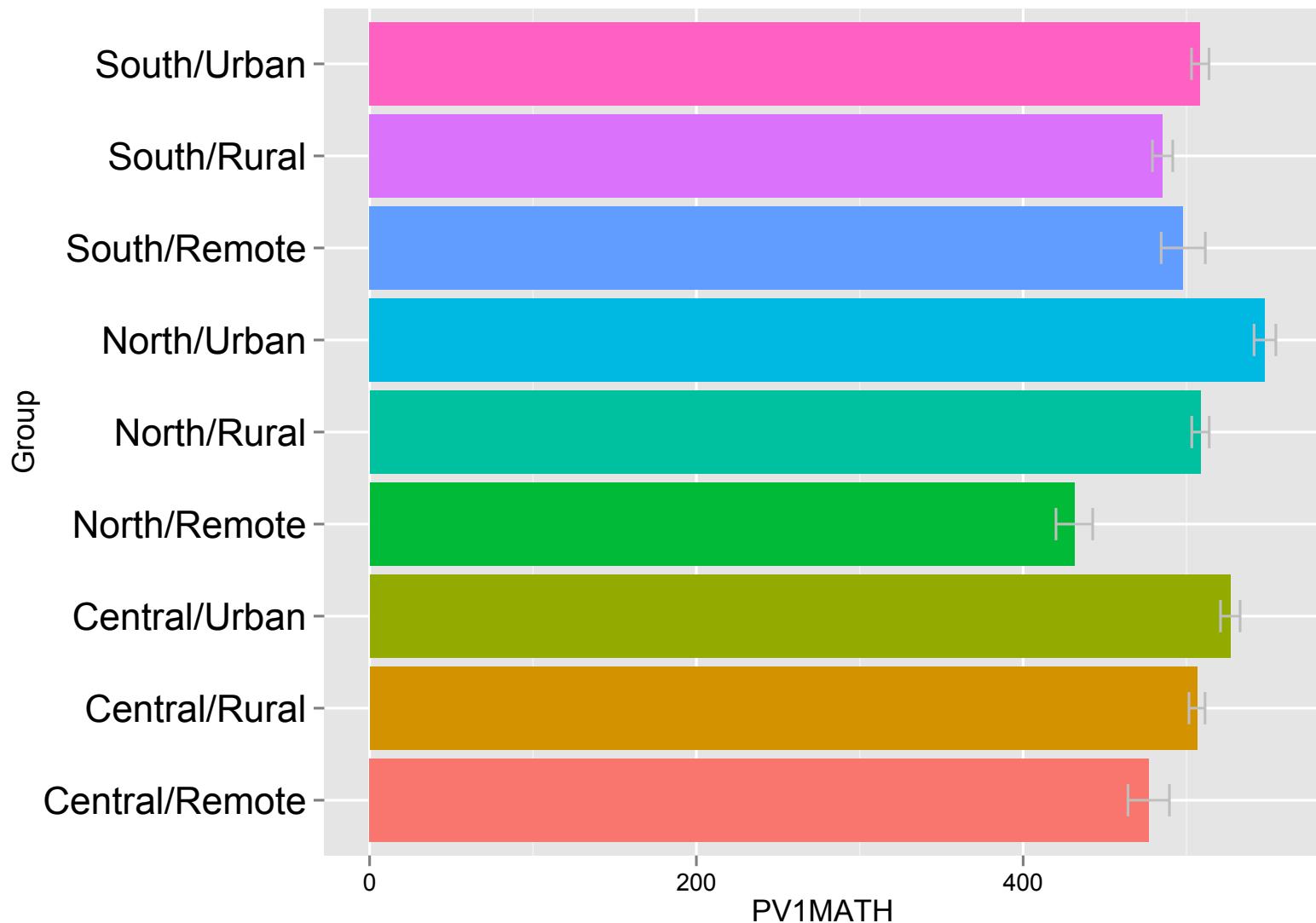


Chuyển trực

```
p = ggplot(pisa, aes(x=Group, y=PVI MATH, fill=Group))  
p = p + stat_summary(fun.y="mean", geom="bar")  
p = p + coord_flip() + theme(legend.position="none")  
p = p + theme(axis.text.x=element_text(angle=0,  
color="black"), axis.text.y=element_text(size=15,  
color="black"))  
p + xlab("") + ylab("")
```



Thêm sai số chuẩn



Thêm số cho từng bar

```
# tính trung bình cho từng nhóm

means = aggregate(pisa$PV1MATH, by=list(pisa$Group),
FUN=mean)

colnames(means) = c("Group", "Mean")

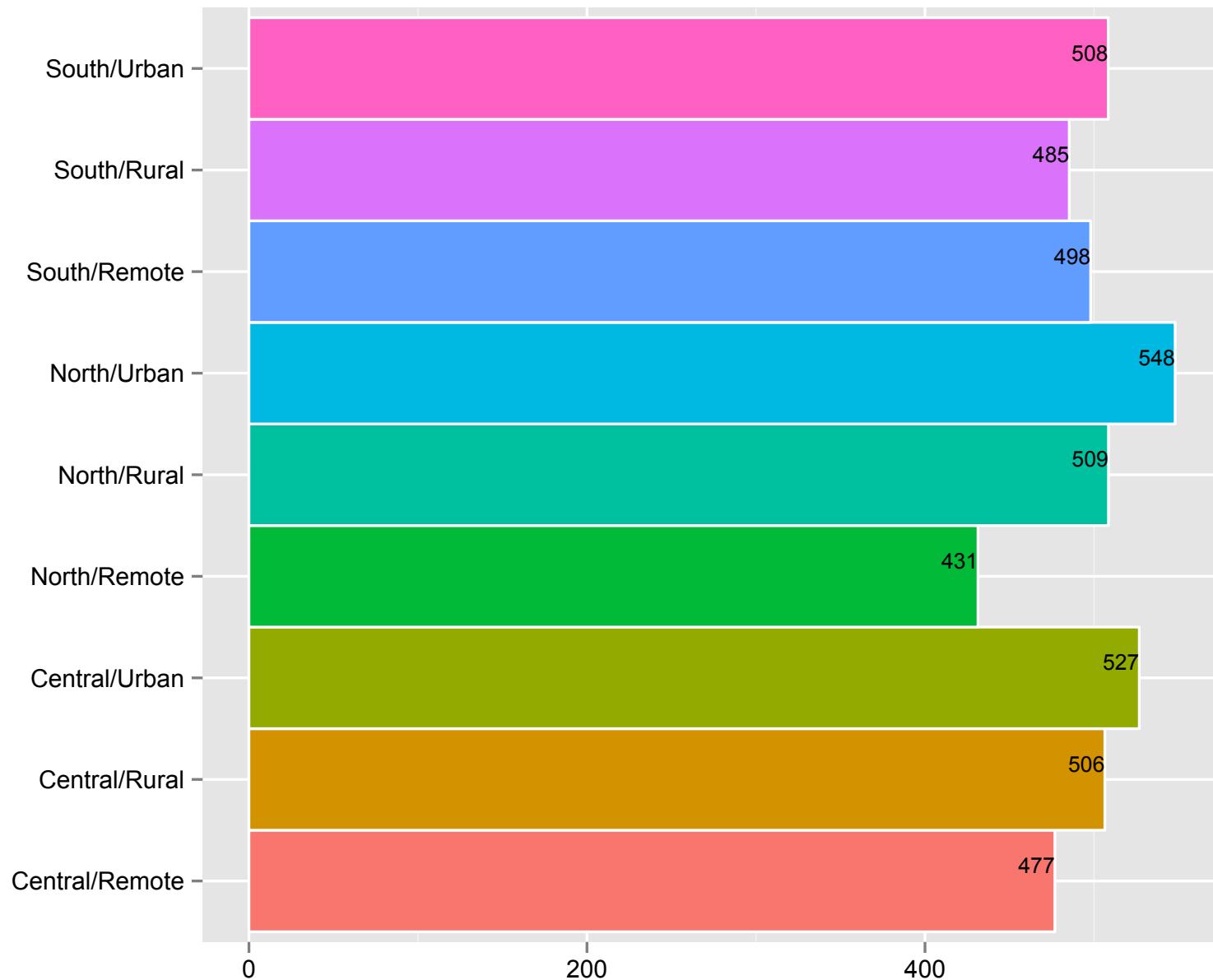
p = ggplot(means, aes(x=Group, y=Mean, fill=Group))

p = p + geom_bar(stat="identity", width=1,
color="white", position=position_dodge())

p = p + theme(legend.position="none") + xlab("") +
ylab("")

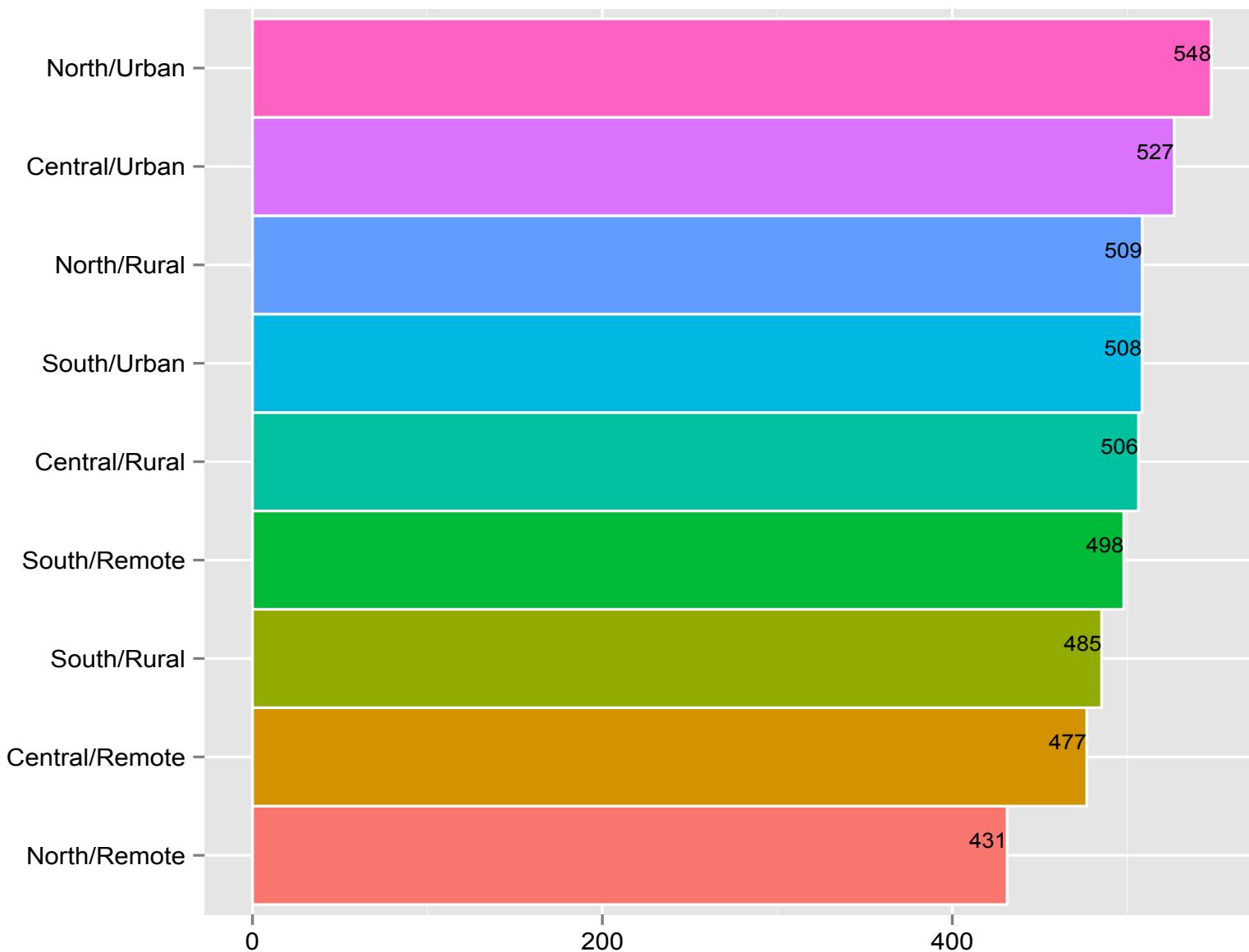
p = p + theme(axis.text.x=element_text(color="black"),
axis.text.y=element_text(color="black"))

p + geom_text(aes(y=Mean, ymax=Mean, label=round(Mean,
0)), position= position_dodge(width=1), size=3,
vjust=-0.5, hjust=1, size=1, color="black") +
coord_flip()
```



Thứ tự theo số trung bình

```
# tính trung bình cho từng nhóm
means = aggregate(pisa$PV1MATH, by=list(pisa$Group),
FUN=mean)
colnames(means) = c("Group", "Mean")
means = transform(means, Group= reorder(Group, Mean))
p = ggplot(means, aes(x=Group, y=Mean, fill=Group))
p = p + geom_bar(stat="identity", width=1,
color="white", position=position_dodge())
p = p + theme(legend.position="none") + xlab("") +
ylab("")
p = p + theme(axis.text.x=element_text(color="black"),
axis.text.y=element_text(color="black"))
p + geom_text(aes(y=Mean, ymax=Mean, label=round(Mean,
0)), position= position_dodge(width=1), size=3,
vjust=-0.5, hjust=1, size=1, color="black") +
coord_flip()
```



**Biểu đồ thanh với số liệu
phân loại (categorical data)**

Dùng barplot trong ggplot2

- Barplot có thể mô tả biến liên tục (trung bình, trung vị, độ lệch chuẩn)
- Barplot cũng có thể dùng để thể hiện biến phân loại, đặc biệt là *frequency* (tần số)

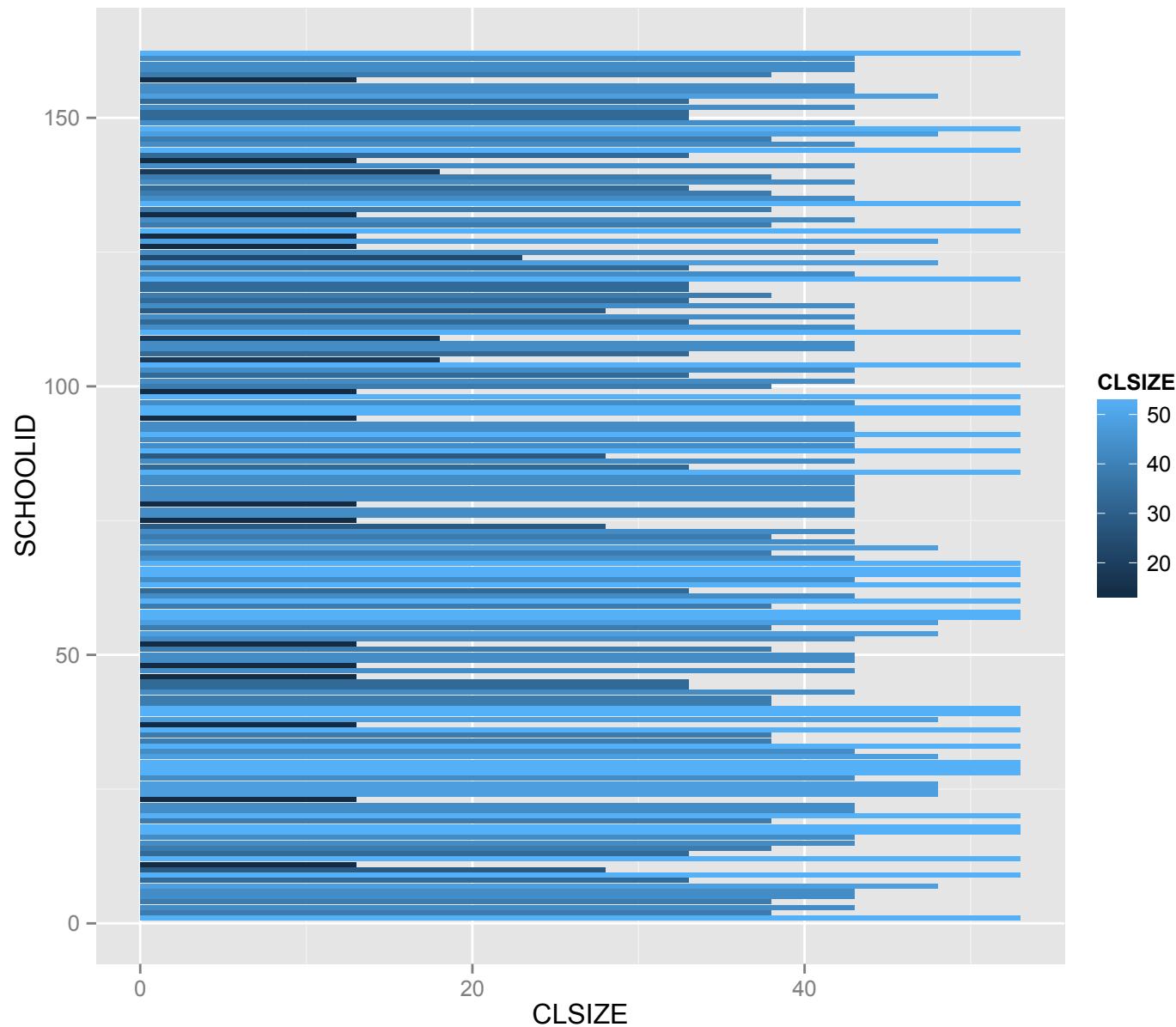
Mô tả từng trường

- Muốn biết CLSIZE của từng trường

```
attach(sc); library(ggplot2)

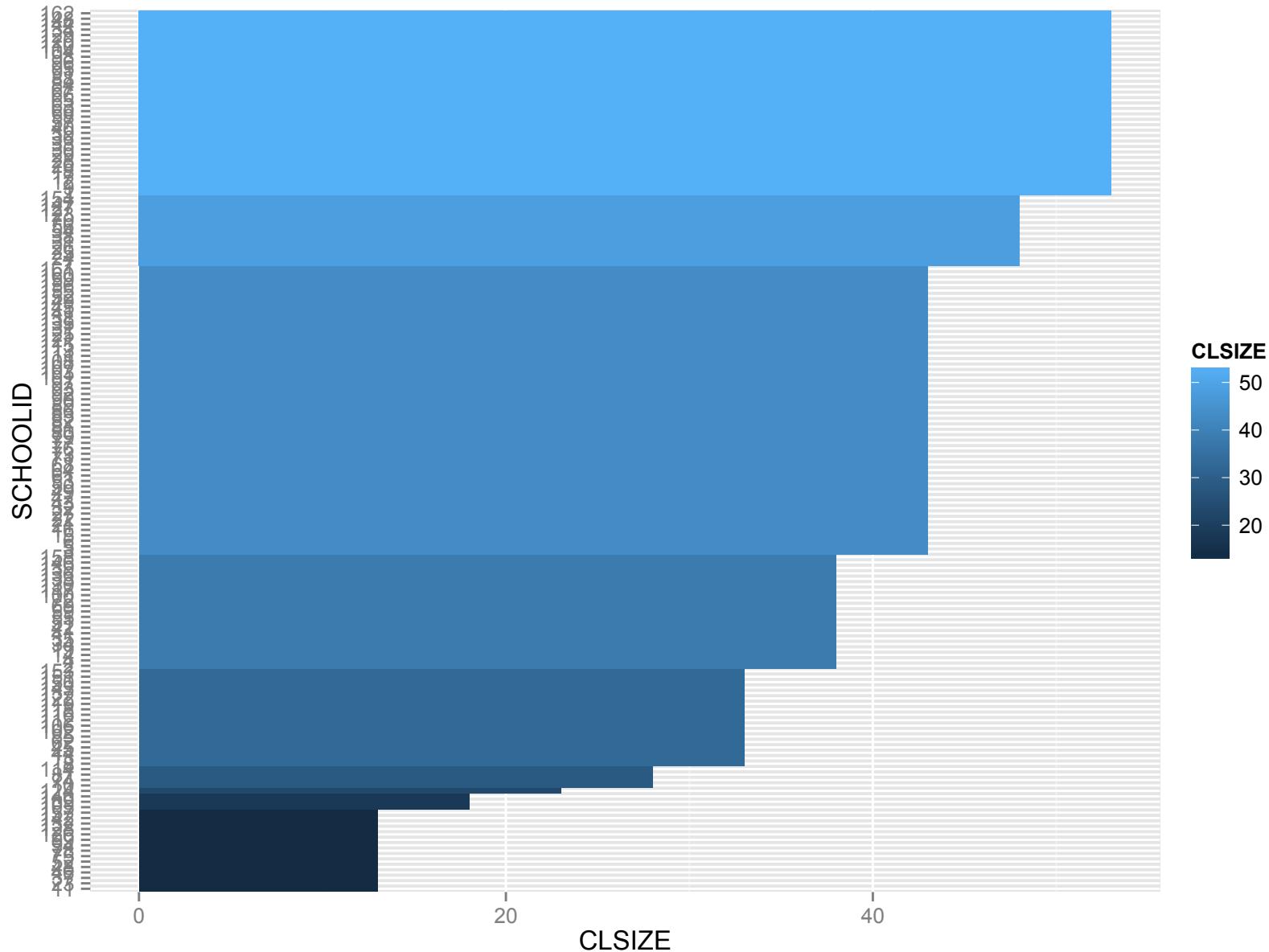
p = ggplot(sc, aes(x=SCHOOLID, y=CLSIZE,
fill=CLSIZE))

p = p + stat_summary(fun.y="mean", geom="bar")
p + coord_flip()
```



Sắp xếp thứ tự ...

```
sc = transform(sc, SCHOOLID=reorder(SCHOOLID,  
CLSIZE) )  
  
p = ggplot(sc, aes(x=SCHOOLID, y=CLSIZE,  
fill=CLSIZE) )  
  
p = p + stat_summary(fun.y="mean", geom="bar")  
p + coord_flip()
```



Phân tích biến phân loại

- Số trường trong từng miền (REGION) và vùng (AREA)

```
p = ggplot(sc, aes(x=REGION, fill=REGION))  
p = p + geom_bar(width=1, colour="white")  
p = p + theme(legend.position="none")  
p = p + coord_flip()  
p + geom_text(stat="bin", color="white", hjust=1,  
size=3, aes(y=..count..,  
label=scales::comma(..count..)))
```

