

Giới thiệu dữ liệu "PISA 2015"

- Kiểm định PISA 2015, Việt Nam
- 5826 học sinh
- Rất nhiều biến số
- File: PISA VN 2015.CSV

```
pisa = read.csv("PISA VN 2015.csv")
```

```
attach(pisa)
```

Đọc file "PISA 2015"

```
pisa = read.csv("~/Dropbox/_Conferences and Workshops/  
Hanoi 2017/PISA VN 2015.csv")  
  
pisa$REGION = factor(pisa$REGION, levels=c("NORTH",  
"CENTRAL", "SOUTH"))  
  
pisa$AREA = factor(pisa$AREA, levels=c("URBAN", "RURAL",  
"REMOTE"))  
  
attach(pisa)
```

```
> head(pisa, 5)
```

	CNTSCHID	CNT	STRATUM	SCHSIZE	CLSIZE	STRATIO	SCHLTYPE	REGION	OECD	AGE
1	70400001	VNM	VNM0313	883	18	22.075	3	SOUTH	0	15.58
2	70400001	VNM	VNM0313	883	18	22.075	3	SOUTH	0	15.92
3	70400001	VNM	VNM0313	883	18	22.075	3	SOUTH	0	15.42
4	70400001	VNM	VNM0313	883	18	22.075	3	SOUTH	0	15.58
5	70400001	VNM	VNM0313	883	18	22.075	3	SOUTH	0	15.92
	ST004D01T	Gender	ISCEDL	ISCEDD	ISCEDO	TEACHSUP	PARED	HEDRES	MISCED	FISCED
1	2	Boys	2	1	1	0.7131	9	-1.0418	1	2
2	2	Boys	2	1	1	0.0322	12	-2.3041	1	4
3	1	Girls	2	1	1	-0.1164	9	-0.7218	1	2
4	1	Girls	2	1	1	-1.1404	5	-1.4595	1	0
5	1	Girls	2	1	1	-0.1128	9	-0.6172	2	2
	HISCED	WEALTH	ESCS	INTBRSCI	INSTSCIE	SCIEEFF	SCIEACT	JOYSCIE	ICTRES	
1	2	-2.0697	-1.7899		NA	0.9798	0.1421	NA	2.1635	-1.5244
2	4	-1.7903	-1.5423		NA	1.7359	-0.8432	NA	2.1635	-1.9305
3	2	-2.1942	-2.0475		NA	-0.2063	-0.1824	NA	-0.1808	-1.6093
4	1	-2.0301	-2.6136		NA	-0.3115	-1.0555	NA	-0.4318	-1.6250
5	2	-1.0522	-1.2179		NA	0.7648	-0.1954	NA	1.3031	-0.5305
	HOMEPOS	HEDRES.1	CULTPOSS	PV1MATH	PV1READ	PV1SCIE	PV1SCEP	PV1SCED	PV1SCID	
1	-2.0537	-1.0418	-0.7273	439.923	412.290	475.612		NA	NA	NA
2	-2.2627	-2.3041	-0.2031	406.251	409.598	450.320		NA	NA	NA
3	-1.9675	-0.7218	-0.2220	414.369	384.307	405.787		NA	NA	NA
4	-2.0686	-1.4595	-0.7039	468.801	459.104	462.968		NA	NA	NA
5	-0.9471	-0.6172	-0.0971	355.432	402.435	453.736		NA	NA	NA

Một số câu hỏi

- Bao nhiêu học sinh từ 3 miền Bắc, Trung, Nam
- Phân bố học sinh vùng sâu xa theo miền
- Phân bố điểm thi môn toán, khoa học và đọc+hiểu
- Phân bố điểm thi môn toán, khoa học và đọc+hiểu theo vùng miền
- Mối tương quan về điểm thi giữa 3 môn

barplot (biểu đồ thanh)

Biểu đồ thanh (bar plot)

Có thể sử dụng để mô tả

- biến số đếm (counts) theo nhóm: bao nhiêu học sinh mỗi vùng?
- biến liên tục (trung bình) theo nhóm: điểm trung bình môn khoa học theo vùng, miền?

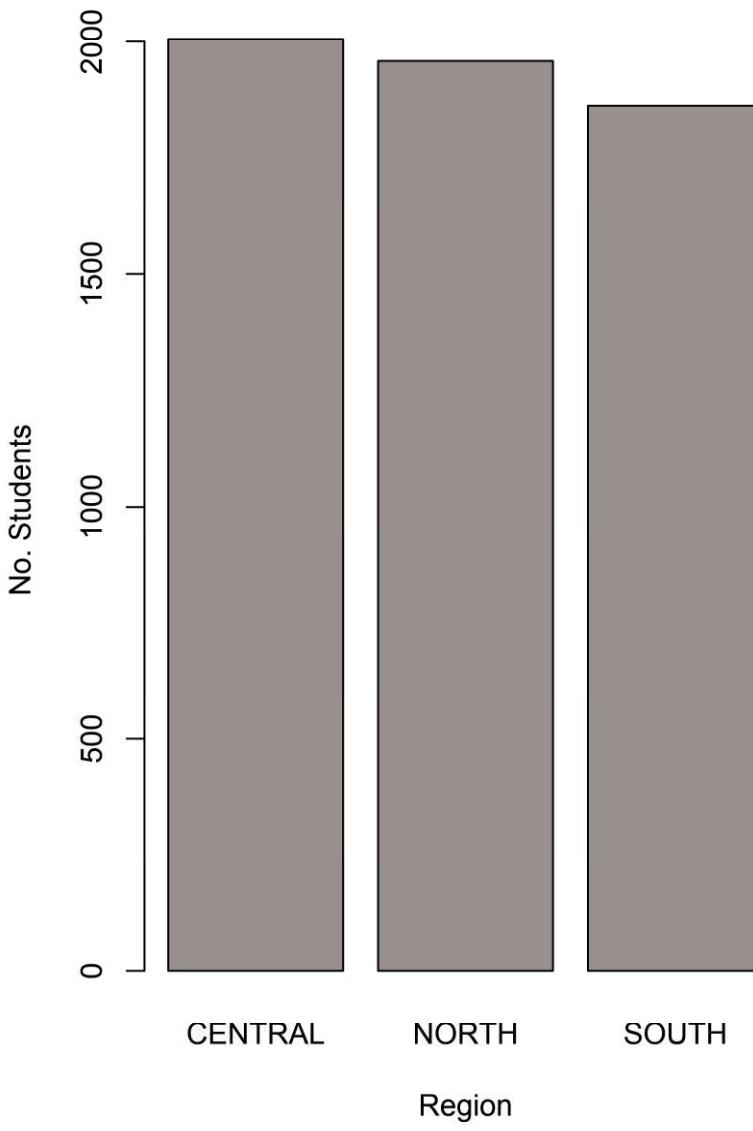
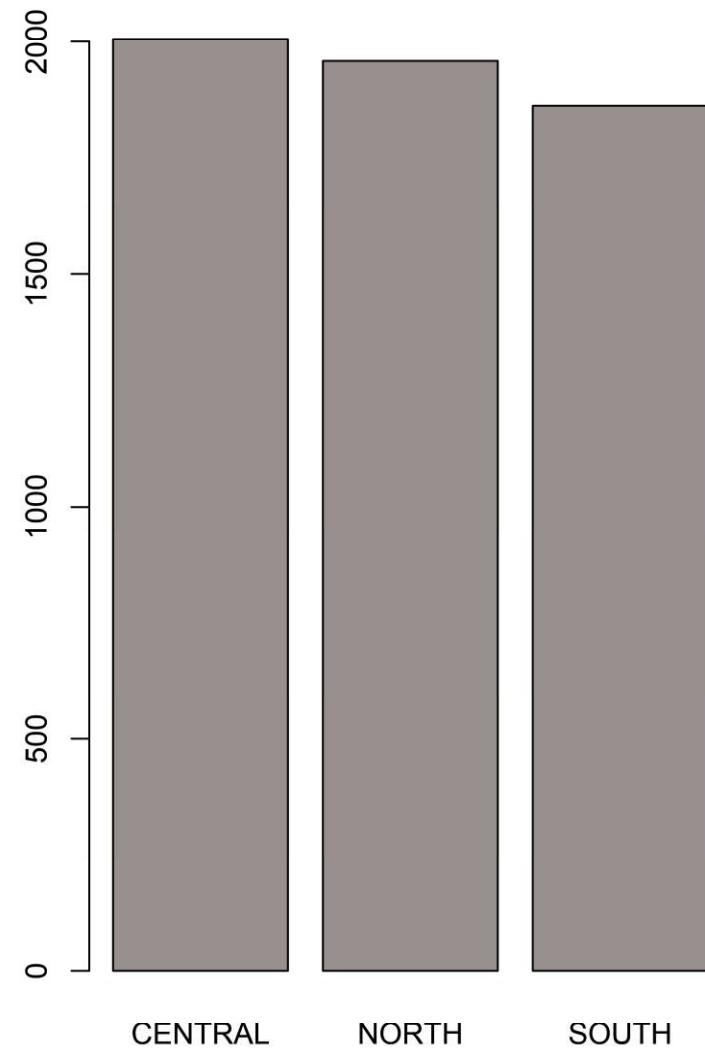
ggplot2 – công cụ lí tưởng cho phân tích

Tần số gãy xương: biểu đồ đơn giản

```
par(mfrow=c(1,2))
```

```
freq = table(REGION)  
barplot(freq)
```

```
barplot(freq, xlab="Region", ylab="No.  
Students")
```



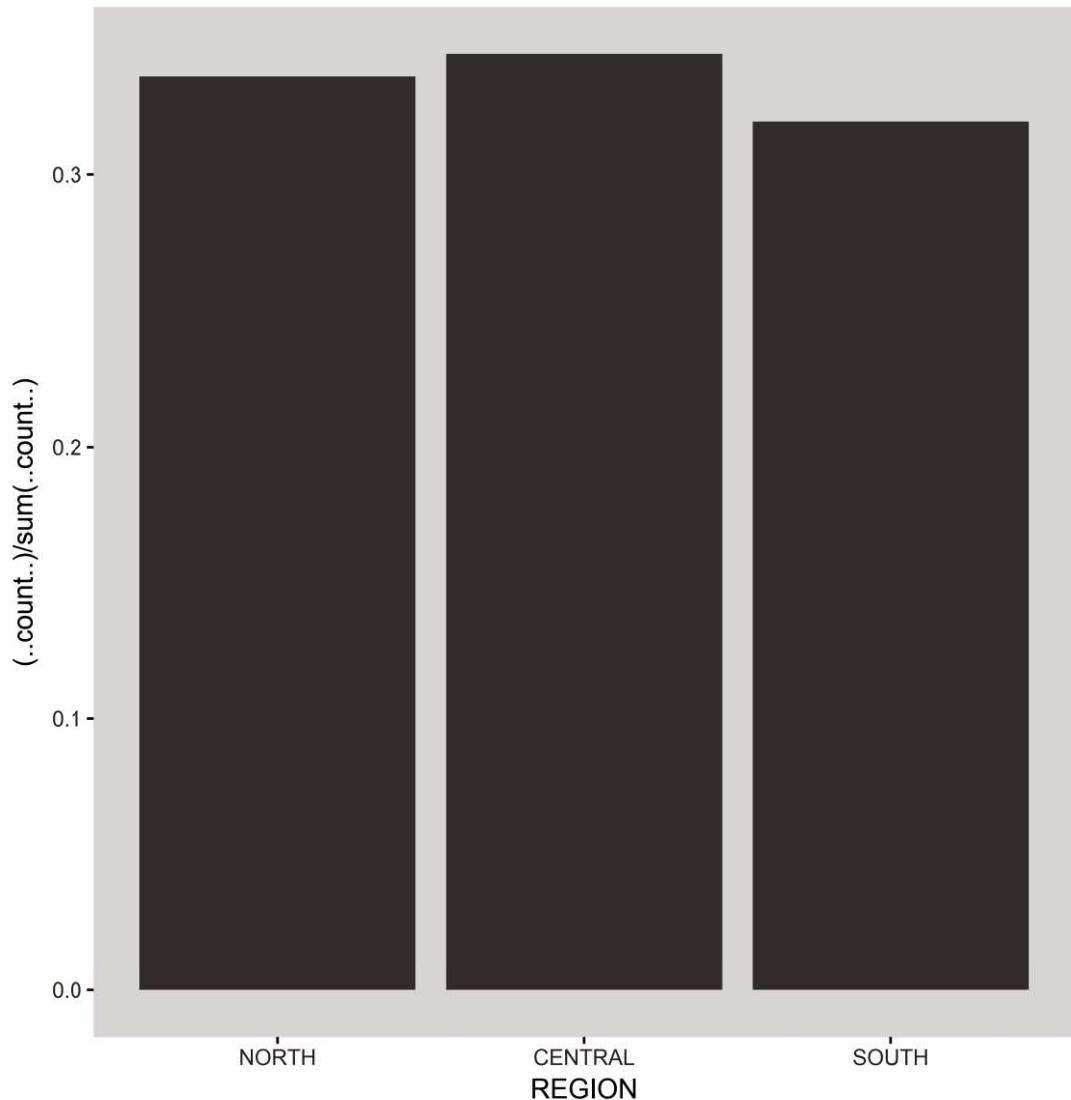
Số học sinh theo vùng – dùng ggplot2

```
library(ggplot2)

p = ggplot(data=pisa,
aes(x=REGION))

p = p + geom_bar(aes(y =
(..count..)/sum(..count..)))

p
```

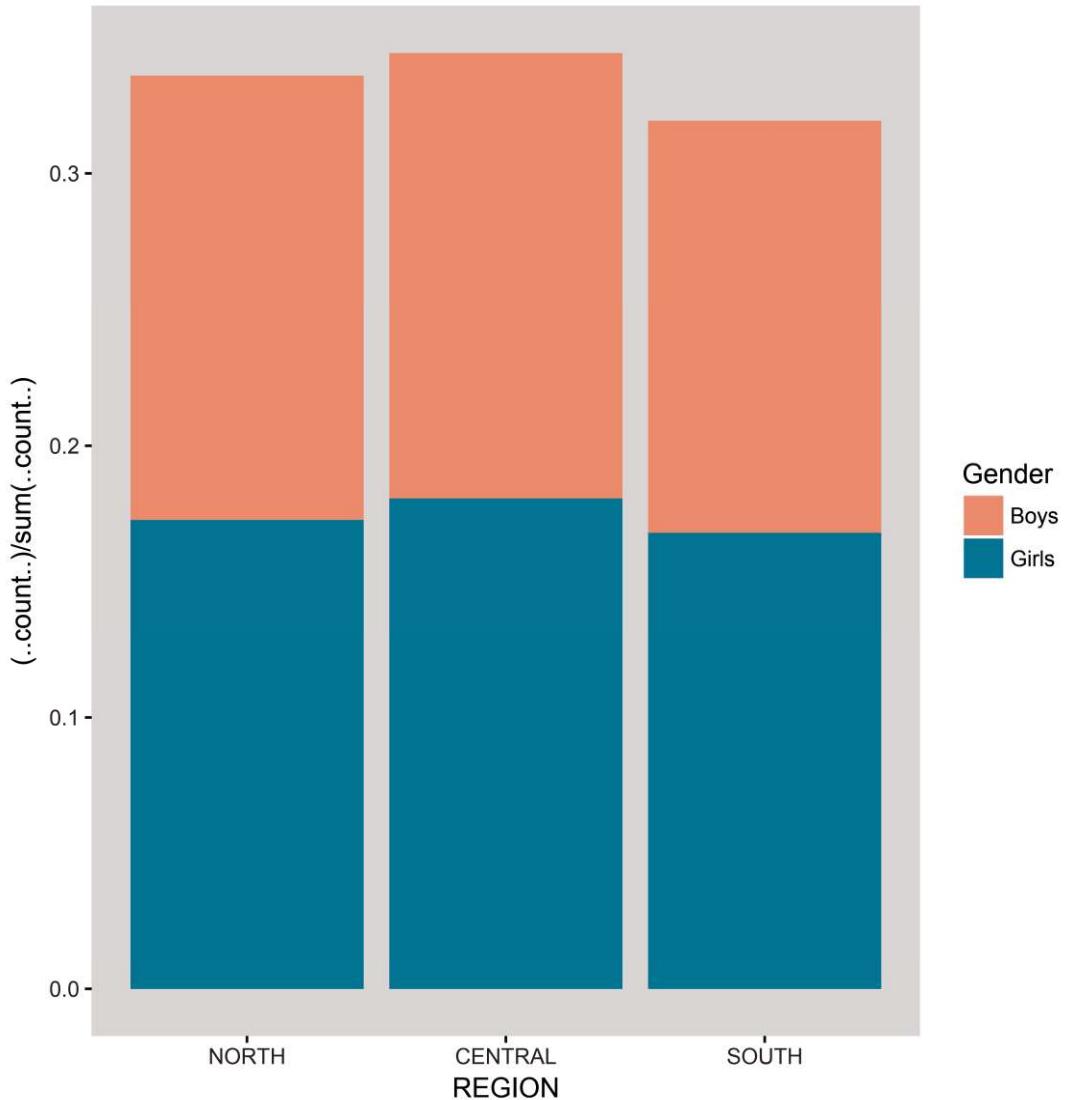


Số học sinh theo vùng và giới tính

```
p = ggplot(data=pisa,  
aes(x=REGION, fill=Gender))
```

```
p = p + geom_bar(aes(y =  
(..count..)/sum(..count..)))
```

```
p
```



Dùng ggplot và phần trăm

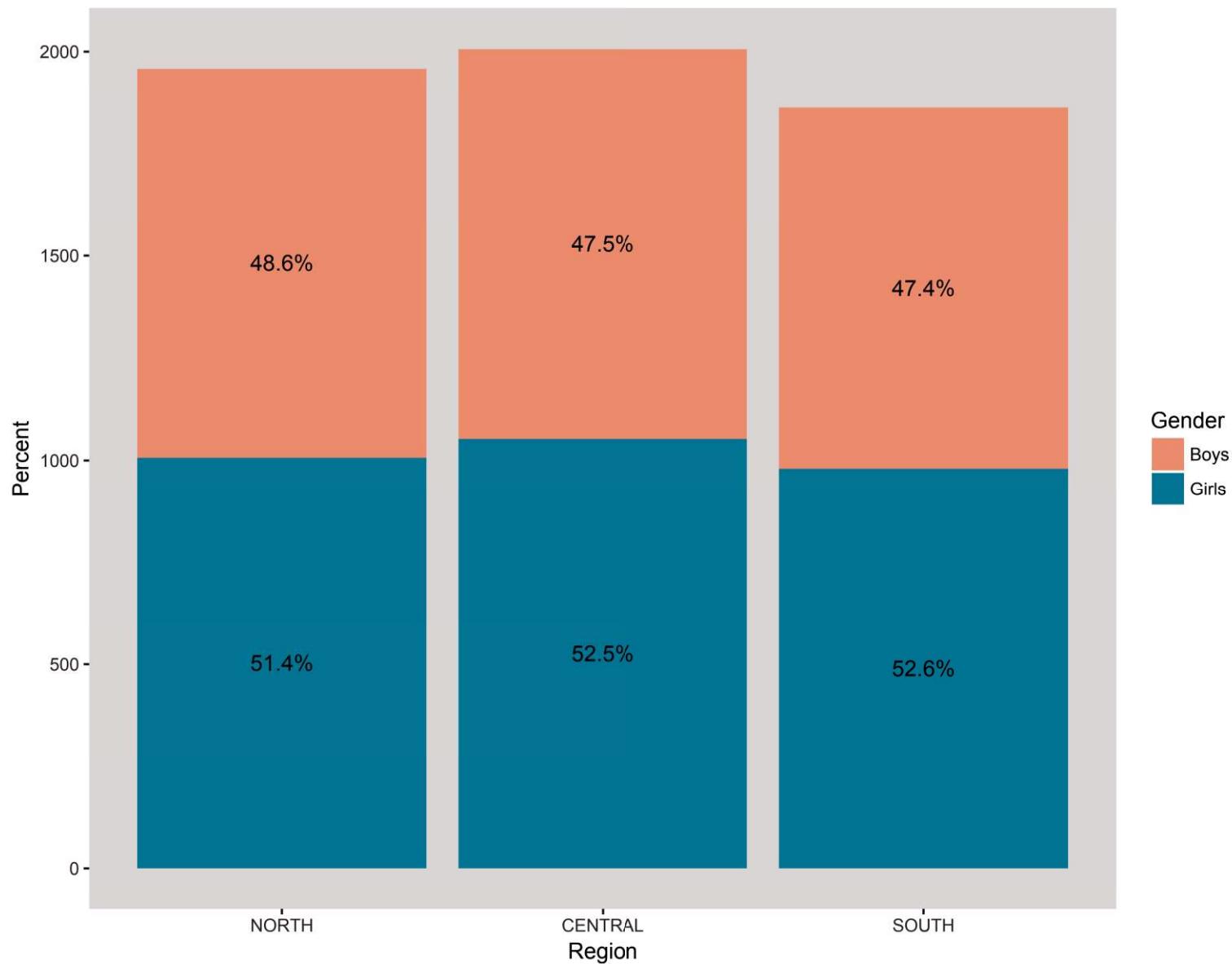
```
library(dplyr)

p = ggplot(pisa %>% count(REGION, Gender) %>%
mutate(pct=n/sum(n)), aes(REGION, n, fill=Gender))

p = p + geom_bar(stat="identity")

p = p + geom_text(aes(label=paste0(sprintf("%1.1f",
pct*100), "%")) , position=position_stack(vjust=0.5))

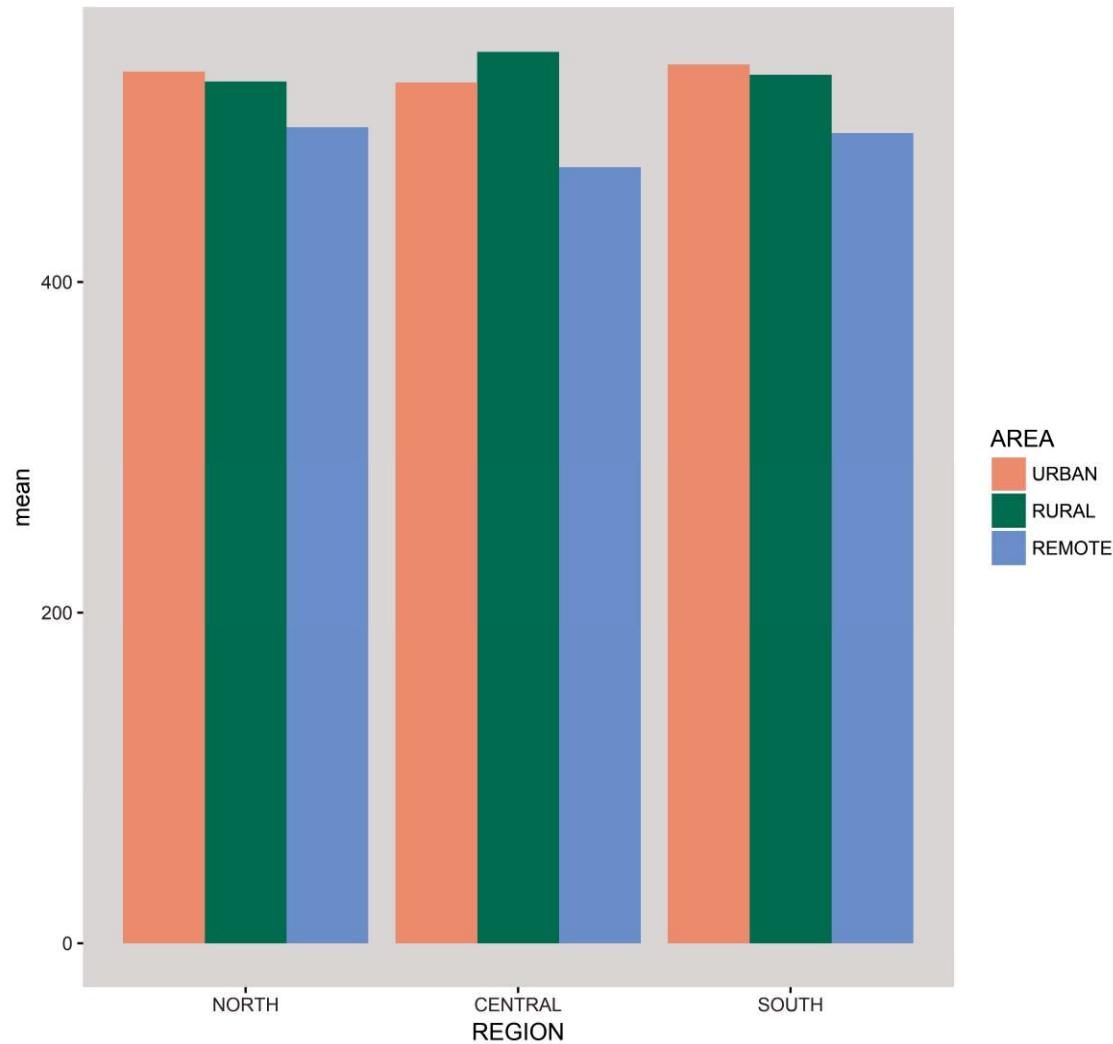
p + xlab("Region") + ylab("Percent")
```



Điểm trung bình môn khoa học theo vùng

```
summary = pisa %>%
  group_by(REGION, AREA) %>%
  summarise(mean =
mean(PV1SCIE, na.rm=T))

p = ggplot(data=summary,
aes(x=REGION, fill=AREA,
y=mean))
p = p + geom_bar(stat =
"identity", position="dodge")
p
```



Điểm trung bình môn khoa học theo HISCED

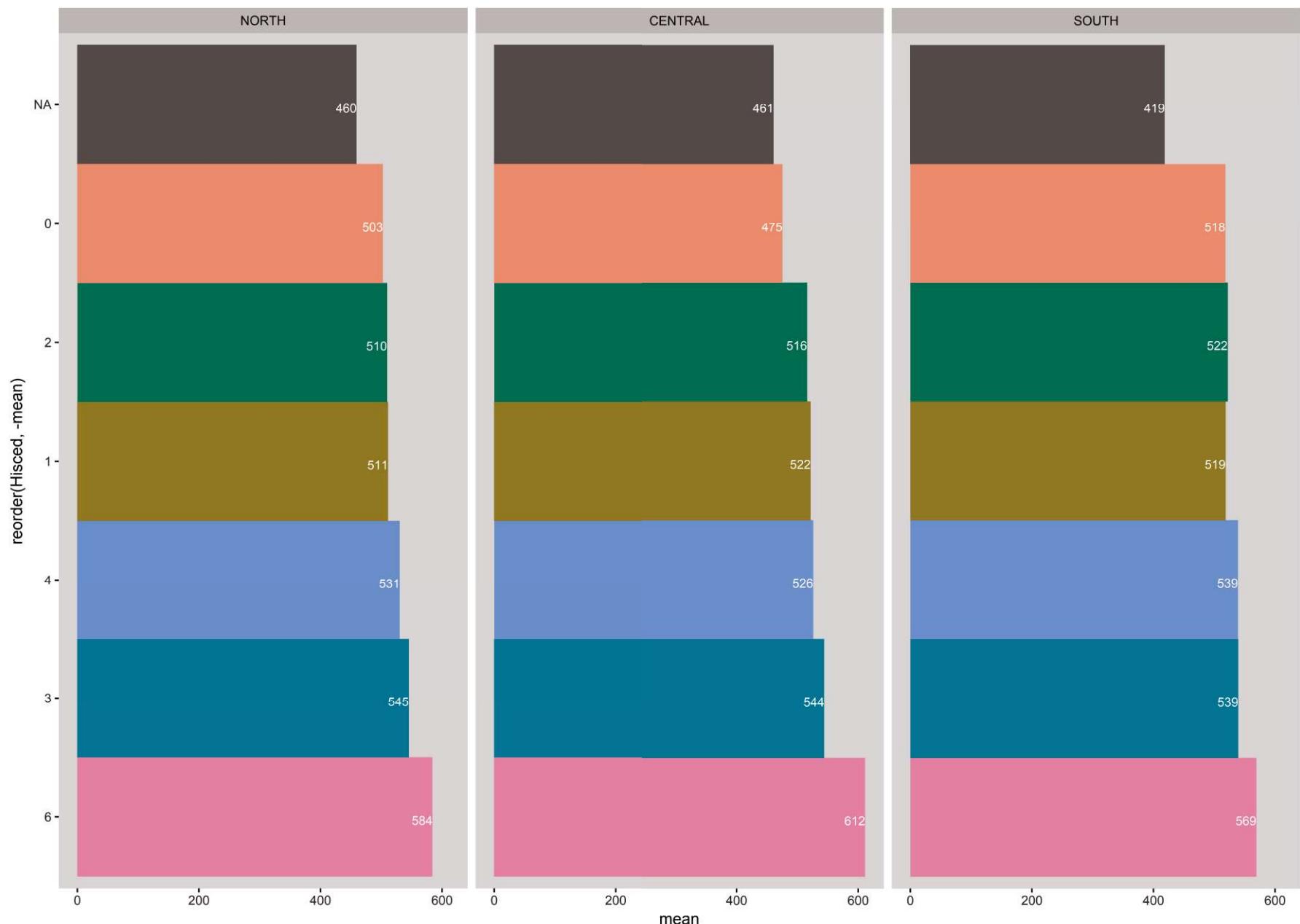
```
pisa$Hisced = as.factor(pisa$HISCED)
attach(pisa)
summary = pisa %>%
  group_by(REGION, Hisced) %>%
  summarise(mean = mean(PV1SCIE, na.rm=T))

p = ggplot(summary, aes(x=reorder(Hisced, -mean), y=mean,
fill=Hisced))

p = p + geom_bar(stat="identity", width=1, position="dodge") +
coord_flip()

p = p + facet_grid(. ~ REGION) + theme(legend.position="none")

p = p + geom_text(aes(y=mean, ymax=mean, label=round(mean, 0)),
position= position_dodge(width=1), size=3, vjust=0.9, hjust=1,
size=3, color="white")
```



Điểm trung bình môn khoa học theo School

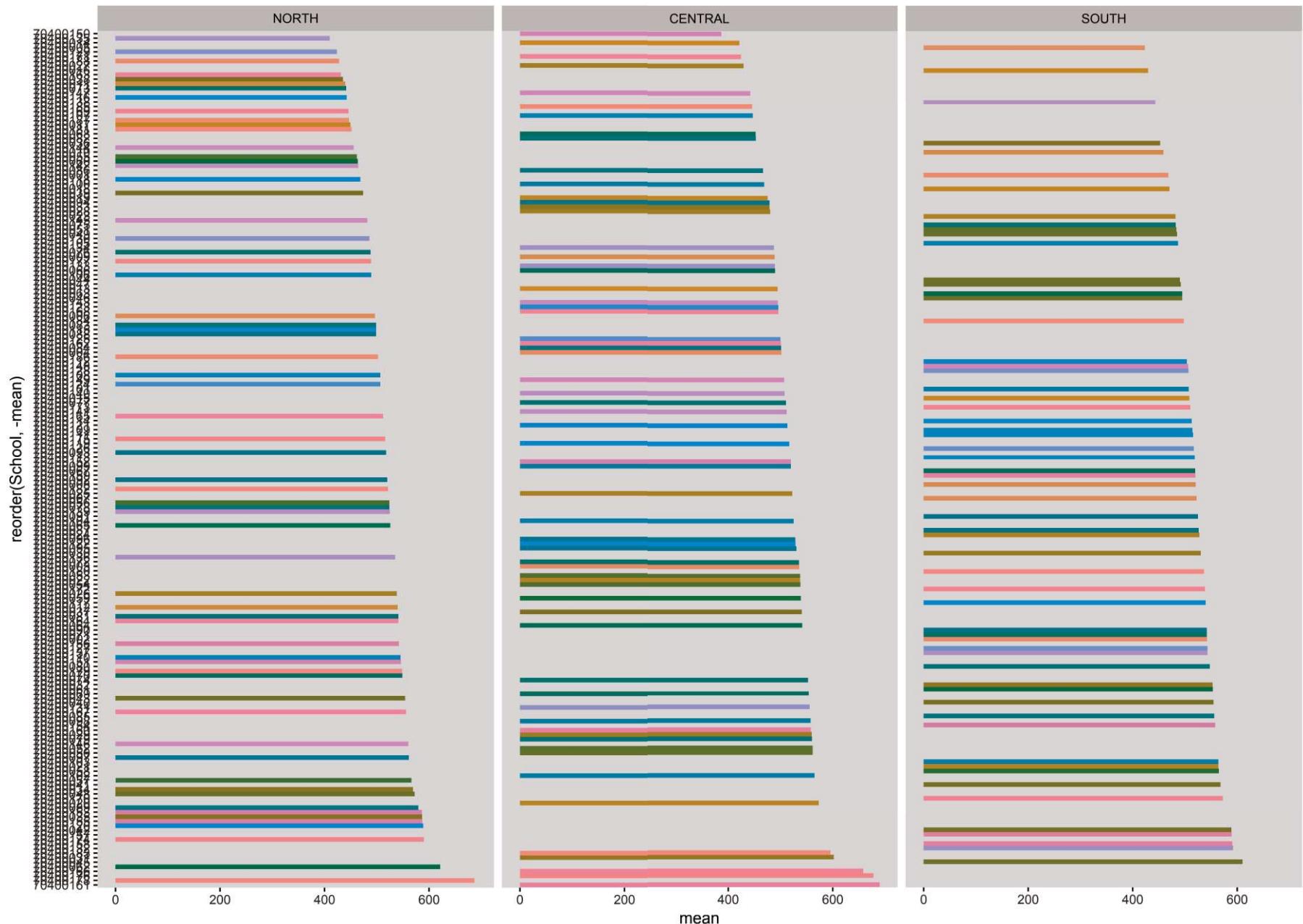
```
pisa$School = as.factor(pisa$CNTSCHID)
attach(pisa)

summary = pisa %>%     group_by(REGION, School) %>%
summarise(mean = mean(PV1SCIE, na.rm=T))

p = ggplot(summary, aes(x=reorder(School, -mean), y=mean, fill=
School))

p = p + geom_bar(stat="identity", width=1, position="dodge") +
coord_flip()

p = p + facet_grid(. ~ REGION) + theme(legend.position="none")
```



Dùng barplot để vẽ số trung bình và sai số chuẩn

```
library(dplyr)

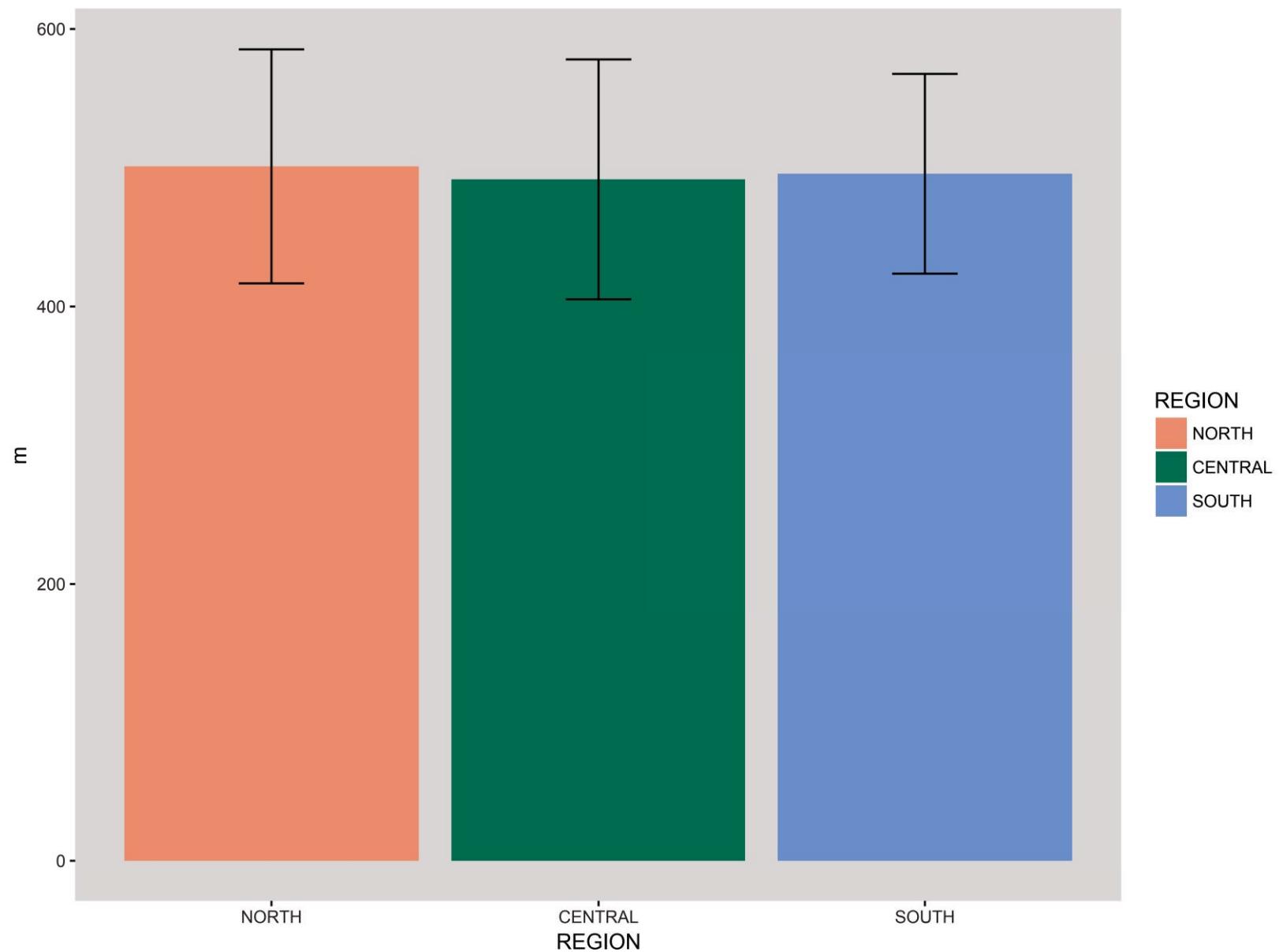
summary = pisa %>% group_by(REGION) %>%
summarise(m = mean(PV1MATH), s = sd(PV1MATH),
nn = n(), se = sd(PV1MATH)/sqrt(n())))

p = ggplot(data=summary, aes(x=REGION, y=m,
fill=REGION))

p = p + geom_bar(stat="identity")

p = p + geom_errorbar(aes(ymin=m-s, ymax=m+s),
width=0.2)

p
```

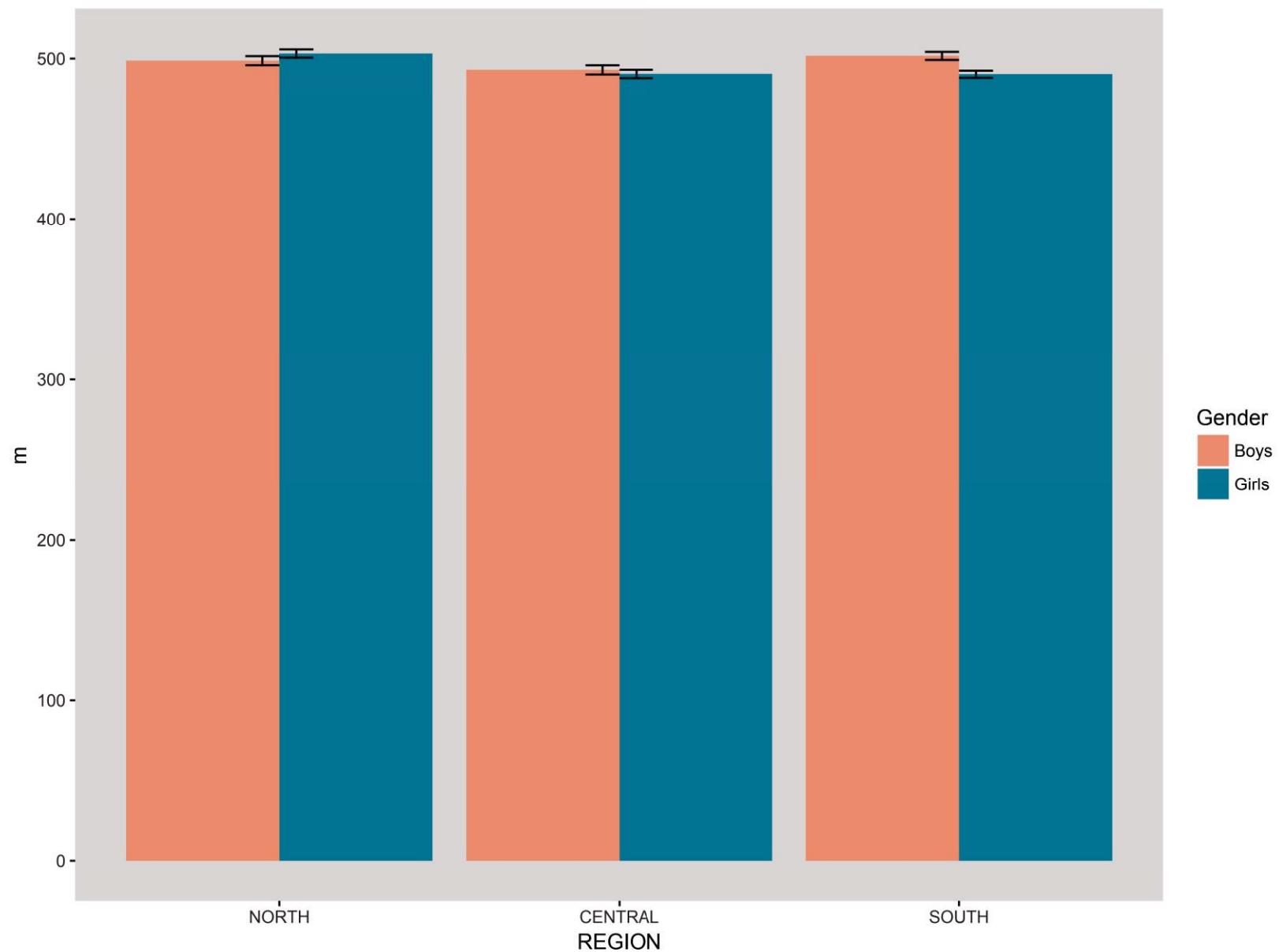


Dùng barplot để vẽ số trung bình và sai số chuẩn

```
summary = pisa %>% group_by(REGION, Gender) %>%
summarise(m = mean(PV1MATH), s = sd(PV1MATH), nn =
n(), se = sd(PV1MATH)/sqrt(n())))

p = ggplot(data=summary, aes(x=REGION, y=m,
fill=Gender))

p = p + geom_bar(stat="identity", position="dodge")
p = p + geom_errorbar(aes(ymin=m-se, ymax=m+se),
position="dodge", width=0.2)
```



Đến phiên bạn ...

Phân bố số học sinh theo vùng và miền

```
library(dplyr)
```

```
p = ggplot(pisa %>% count(REGION, AREA) %>%
mutate(pct=n/sum(n)), aes(REGION, n, fill=AREA))

p = p + geom_bar(stat="identity")

p = p + geom_text(aes(label=paste0(sprintf("%1.1f",
pct*100), "%")), position=position_stack(vjust=0.5))

p + xlab("Region") + ylab("Percent")
```

Phân bố HISCED theo giới tính

```
# tạo biến factor
```

```
pisa$HISCED = as.factor(pisa$HISCED)
```

```
pisa$PARED = as.factor(pisa$PARED)
```

```
attach(pisa)
```

```
# vẽ biểu đồ
```

```
p = ggplot(data=pisa, aes(x=HISCED, fill=Gender))
```

```
p = p + geom_bar(aes(y = ..count../sum(..count..)))
```

```
p + xlab("Parental Highest Education")
```

Điểm trung bình môn khoa học theo vùng

```
p = ggplot(pisa, aes(x=REGION, y=PV1SCIE,  
fill=REGION))
```

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

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

Điểm trung bình cho môi trường miền Trung

```
central = subset(pisa, REGION=="CENTRAL")

means = aggregate(central$PV1MATH, by=list(central$School),
FUN=mean)

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

mean = transform(means, School=reorder(School, order(Mean,
decreasing=T)))

p = ggplot(mean, aes(x=reorder(School, Mean), y=Mean, fill=School))

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

p = p + coord_flip() + theme(legend.position="none")

p + geom_text(aes(y=Mean, ymax=Mean, label=round(Mean, 0)),
position= position_dodge(width=1), size=2, vjust=0.9, hjust=1,
size=2, color="white")
```

histogram (biểu đồ phân bố)

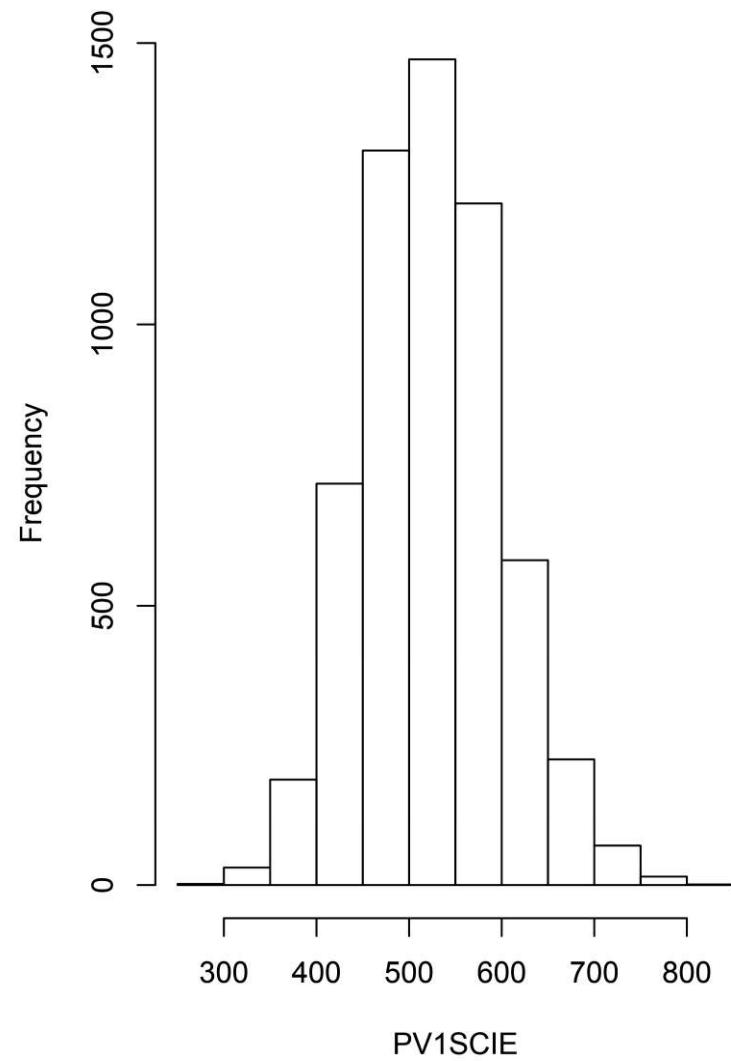
Phân bố điểm môn khoa học

```
par(mfrow=c(1,2))
```

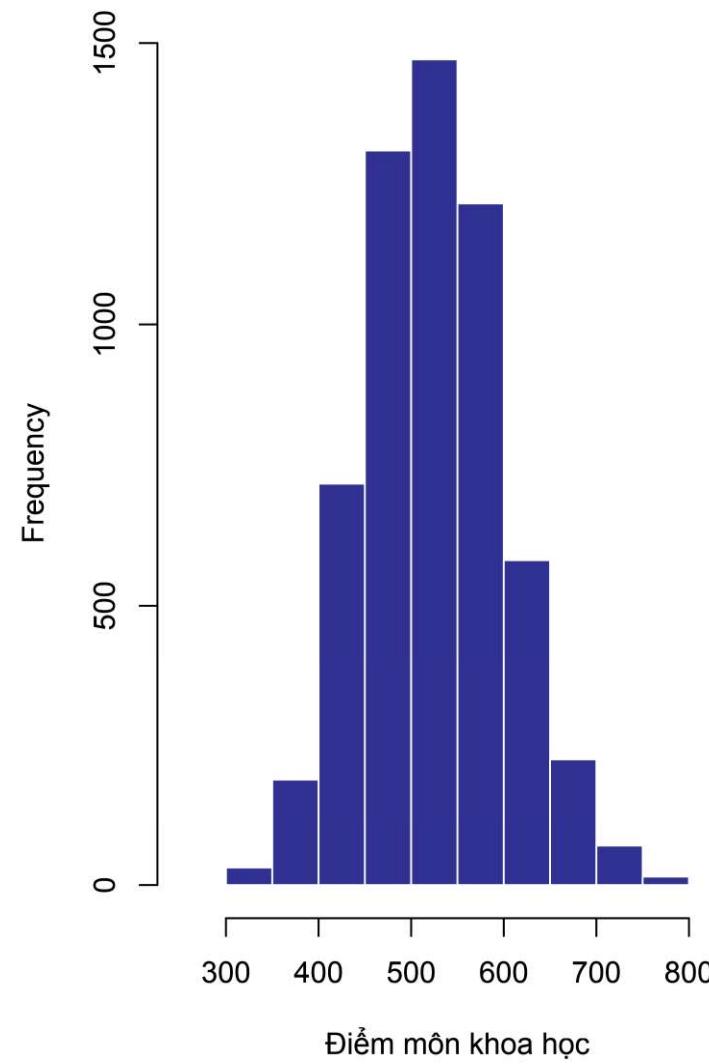
```
hist(PV1SCIE)
```

```
hist(PV1SCIE, col="blue", border="white", xlab="Điểm  
môn khoa học", main="Phân bố điểm môn khoa học")
```

Histogram of PV1SCIE



Phân bố điểm môn khoa học



Phân bố điểm môn khoa học: thêm "density"

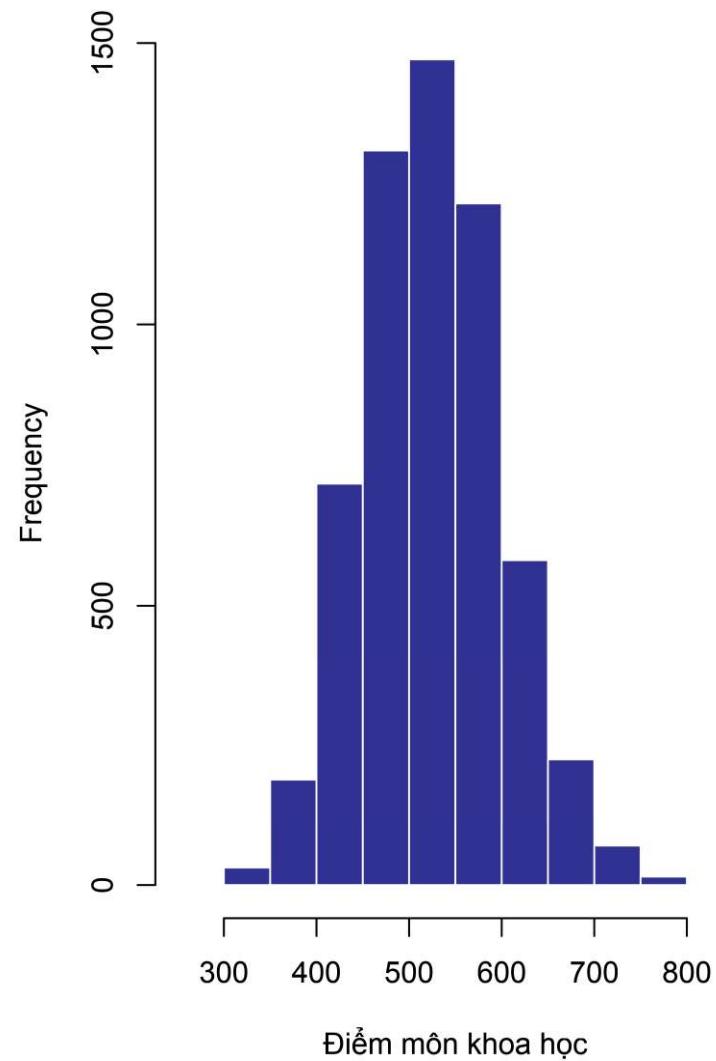
```
par(mfrow=c(1,2))

hist(PV1SCIE, col="blue", border="white", xlab="Điểm  
môn khoa học", main="Phân bố điểm môn khoa học")

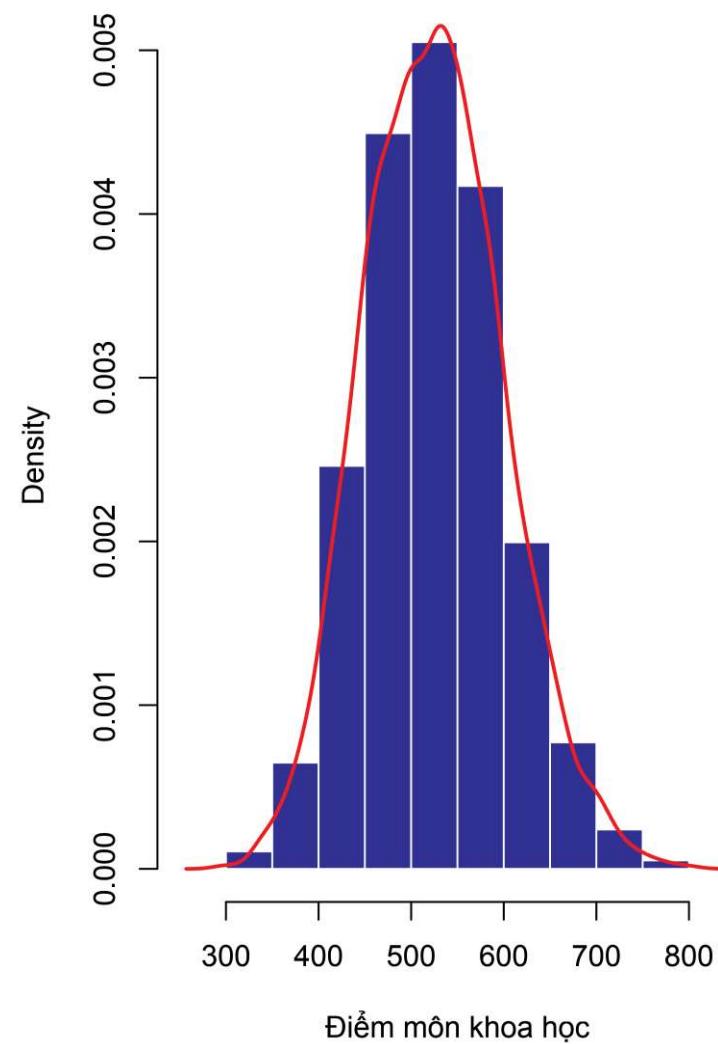
# thêm density line
hist(PV1SCIE, prob=T, col="blue", border="white",
xlab="Điểm môn khoa học", main="Phân bố điểm môn  
khoa học")

lines(density(PV1SCIE, na.rm=T), col="red", lwd=2)
```

Phân bố điểm môn khoa học

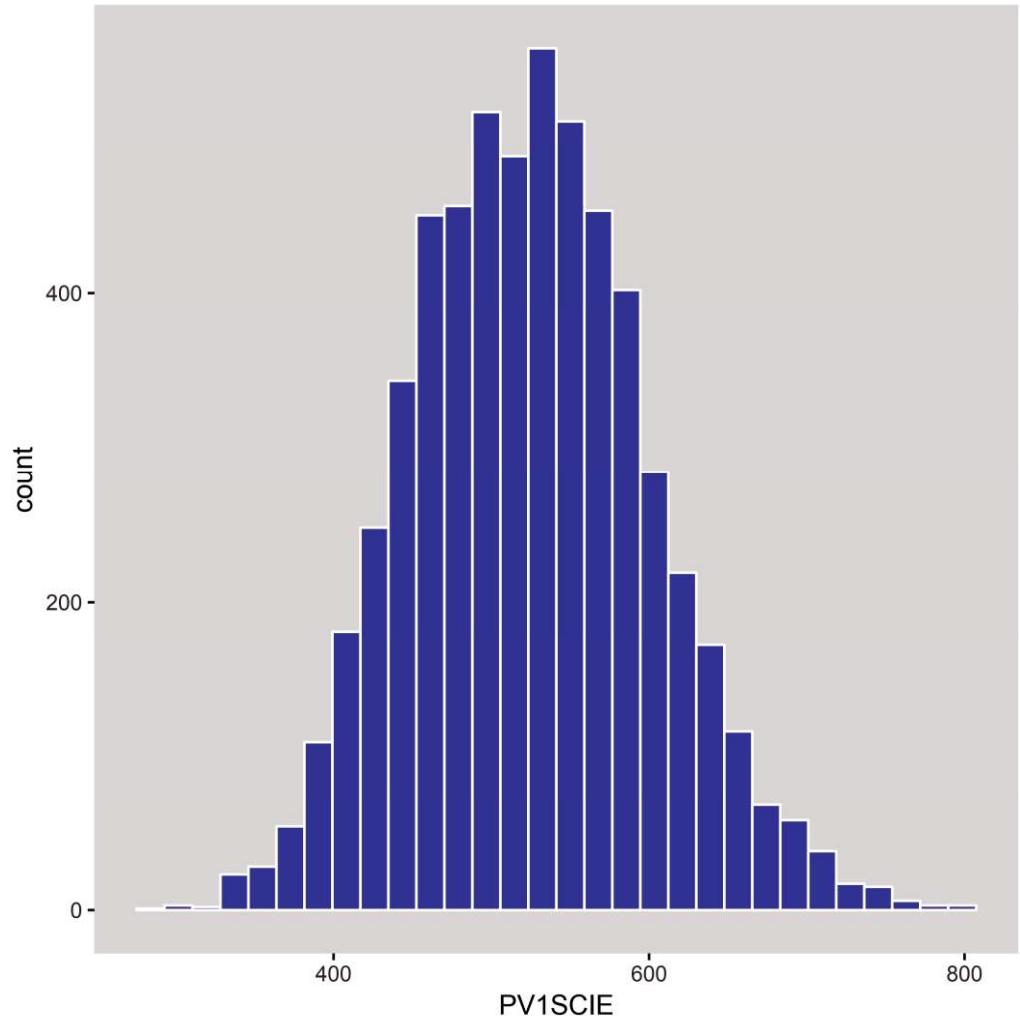


Phân bố điểm môn khoa học



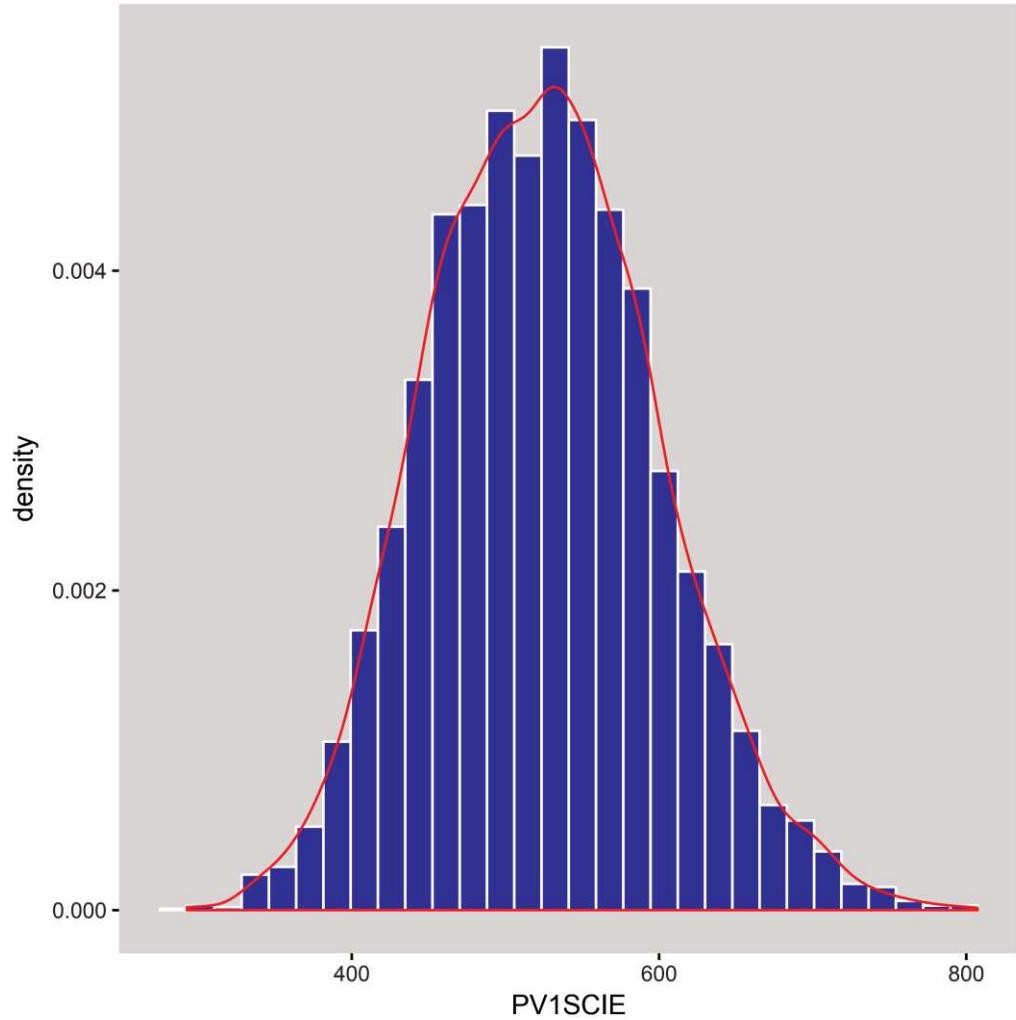
Biểu đồ phân bố dùng **ggplot2**: đơn giản

```
library(ggplot2);  
library(gridExtra)  
  
p = ggplot(pisa,  
aes(x=PV1SCIE))  
  
p = p +  
geom_histogram(color="white"  
, fill="blue")
```



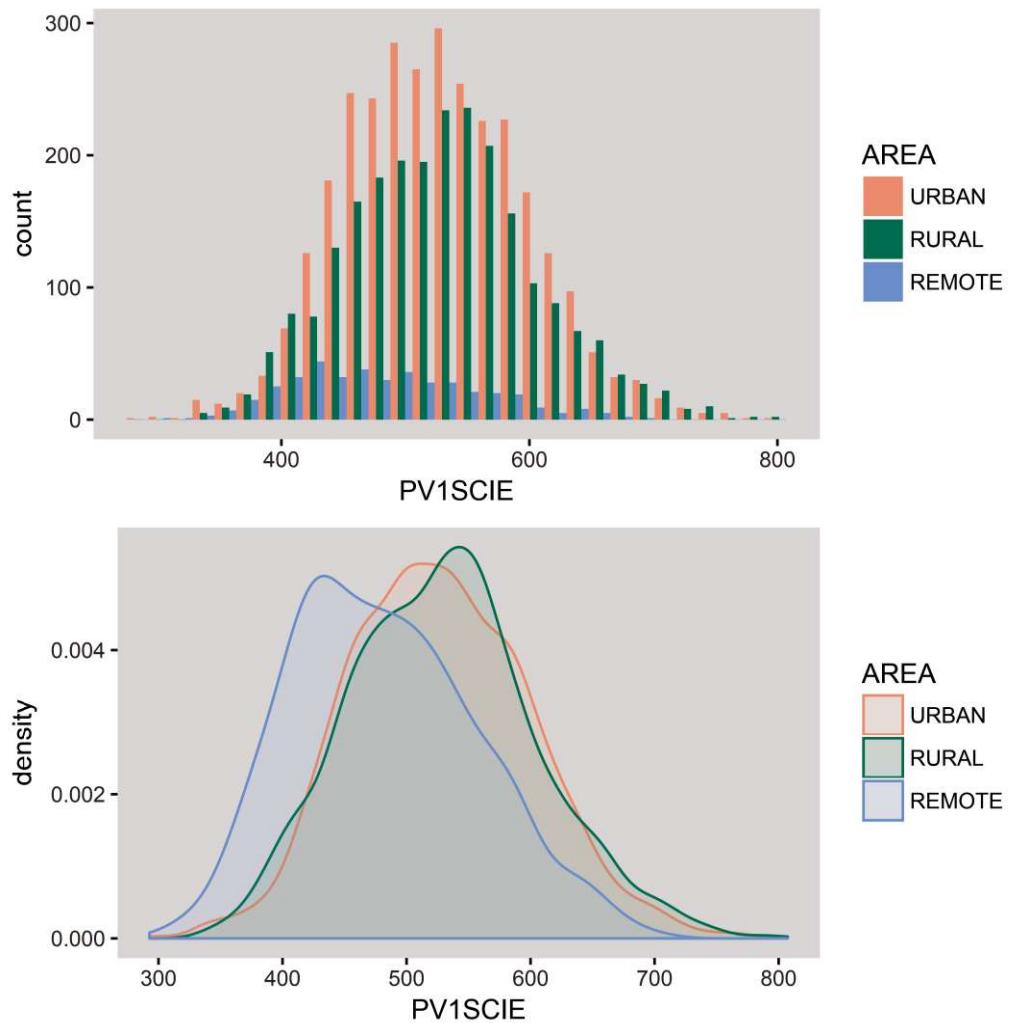
Biểu đồ phân bố dùng **ggplot2**: đơn giản + **density**

```
p = ggplot(pisa,  
aes(x=PV1SCIE))  
  
p = p +  
geom_histogram(aes(y=..density..), color="white",  
fill="blue")  
  
p = p +  
geom_density(col="red")
```



Dùng ggplot2: đơn giản + density + theo nhóm

```
p = ggplot(pisa,  
aes(x=PV1SCIE, fill=AREA))  
  
p1 = p +  
geom_histogram(position="dodge")  
  
p2 = ggplot(pisa,  
aes(x=PV1SCIE, fill=AREA,  
color=AREA)) +  
geom_density(alpha = 0.1)  
  
grid.arrange(p1, p2, nrow=2)
```



Đến phiên bạn ...

```
library(ggplot2); library(gridExtra)
p = ggplot(pisa, aes(x=PV1READ, fill=REGION))
p1 = p + geom_histogram(position="dodge")

p2 = ggplot(pisa, aes(x=PV1READ, fill=REGION,
color=REGION)) + geom_density(alpha = 0.1)

grid.arrange(p1, p2, nrow=2)
```

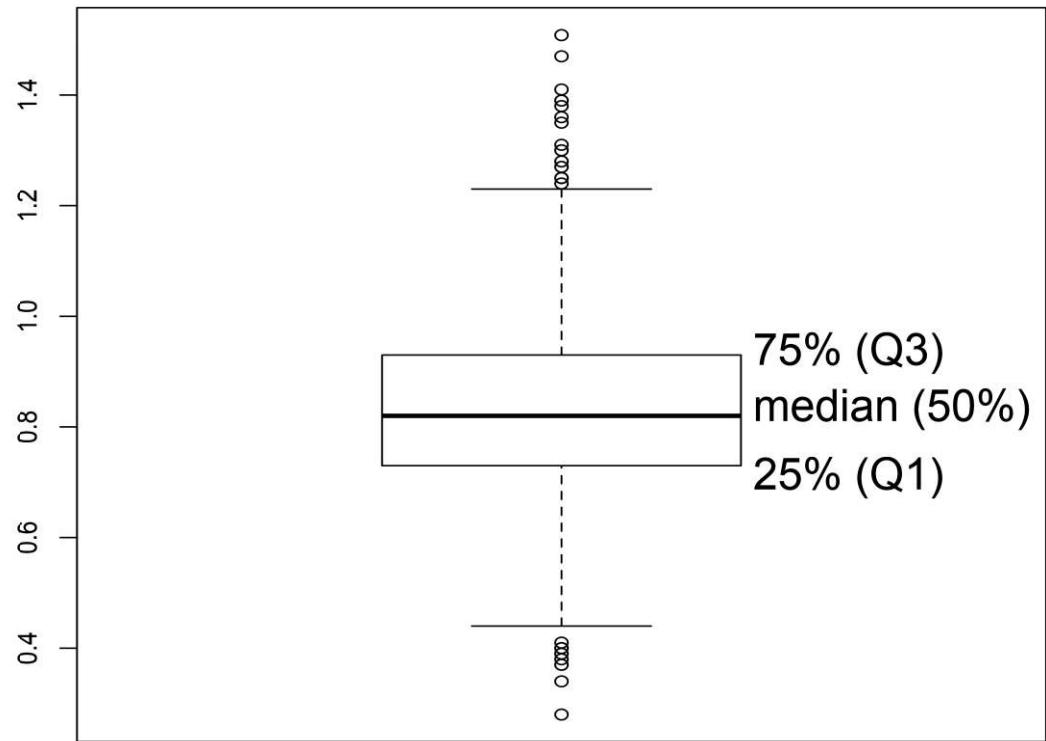
Đến phiên bạn ...

```
p = ggplot(pisa, aes(x=PV1READ, fill=AREA))  
p1 = p + geom_histogram(position="dodge") +  
facet_grid(REGION ~ .)  
  
p2 = p + geom_density(aes(x=PV1READ, color=Gender,  
alpha=0.5)) + facet_wrap(~Gender)  
  
grid.arrange(p1, p2, nrow=2)
```

boxplot (biểu đồ hộp)

5 yếu tố trong biểu đồ hộp

- median = trung vị
- 2 hinges = 25% và 75% bách phân
- fences = $1.5 \times$ interquartile range
- whiskers = nối hai hinges
- Outliers (có thể giá trị ngoại vi)



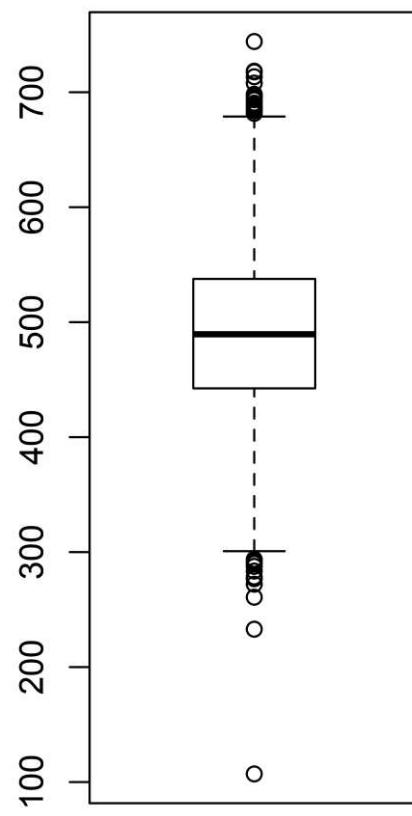
interquartile range = Q3 – Q1

Điểm môn đọc & hiểu

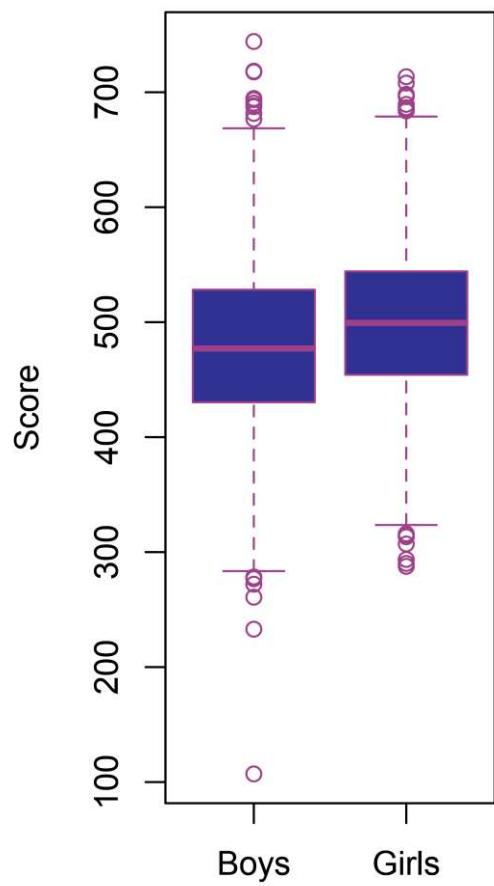
```
par(mfrow=c(1, 2))
```

```
boxplot(PV1READ)
```

```
boxplot(PV1READ ~ Gender,  
main="Phân bố điểm môn đọc &  
hiểu", xlab="Điểm môn đọc &  
hiểu", ylab="Score",  
col="blue", border="purple")
```



Phân bố điểm môn đọc & hiểu



Điểm môn đọc & hiểu

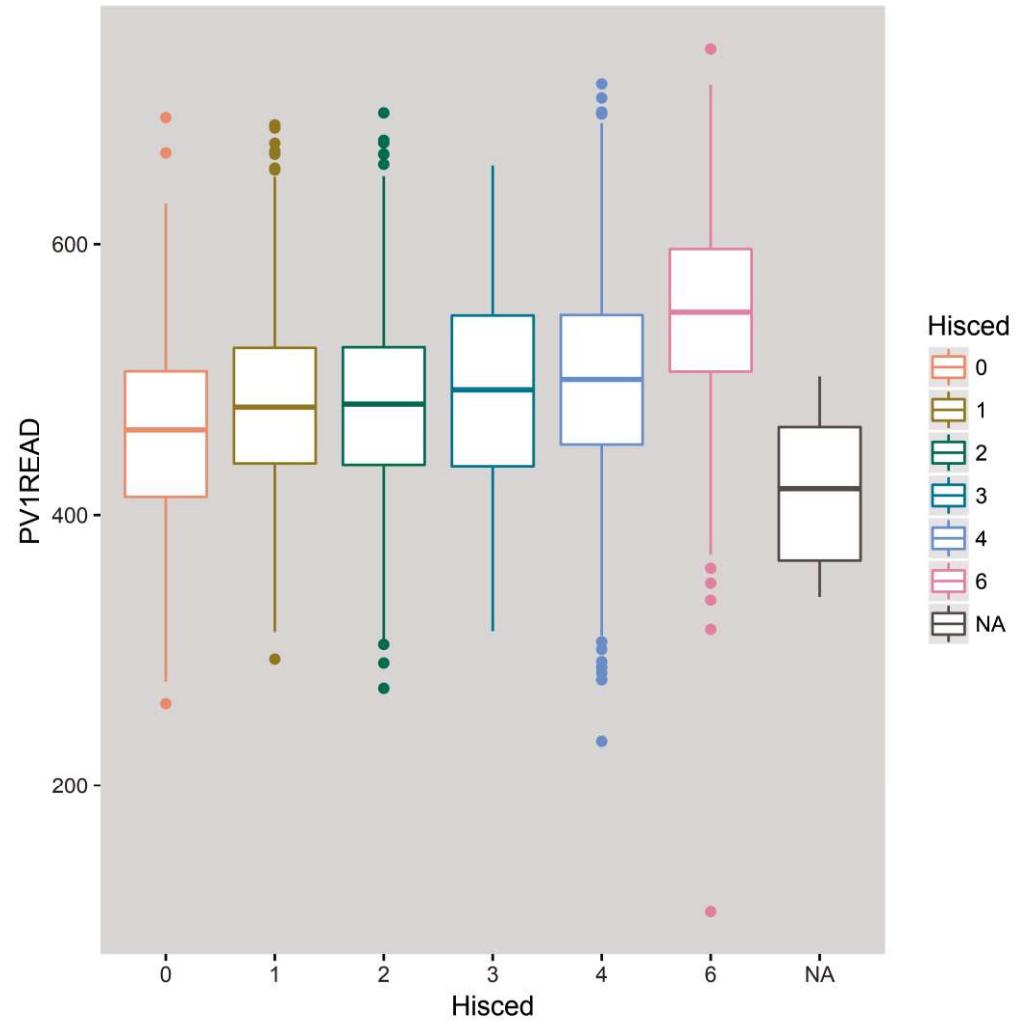
Biểu đồ hộp dùng ggplot2: đơn giản

```
# tạo biến factor
```

```
Hisced = as.factor(HISCED)
```

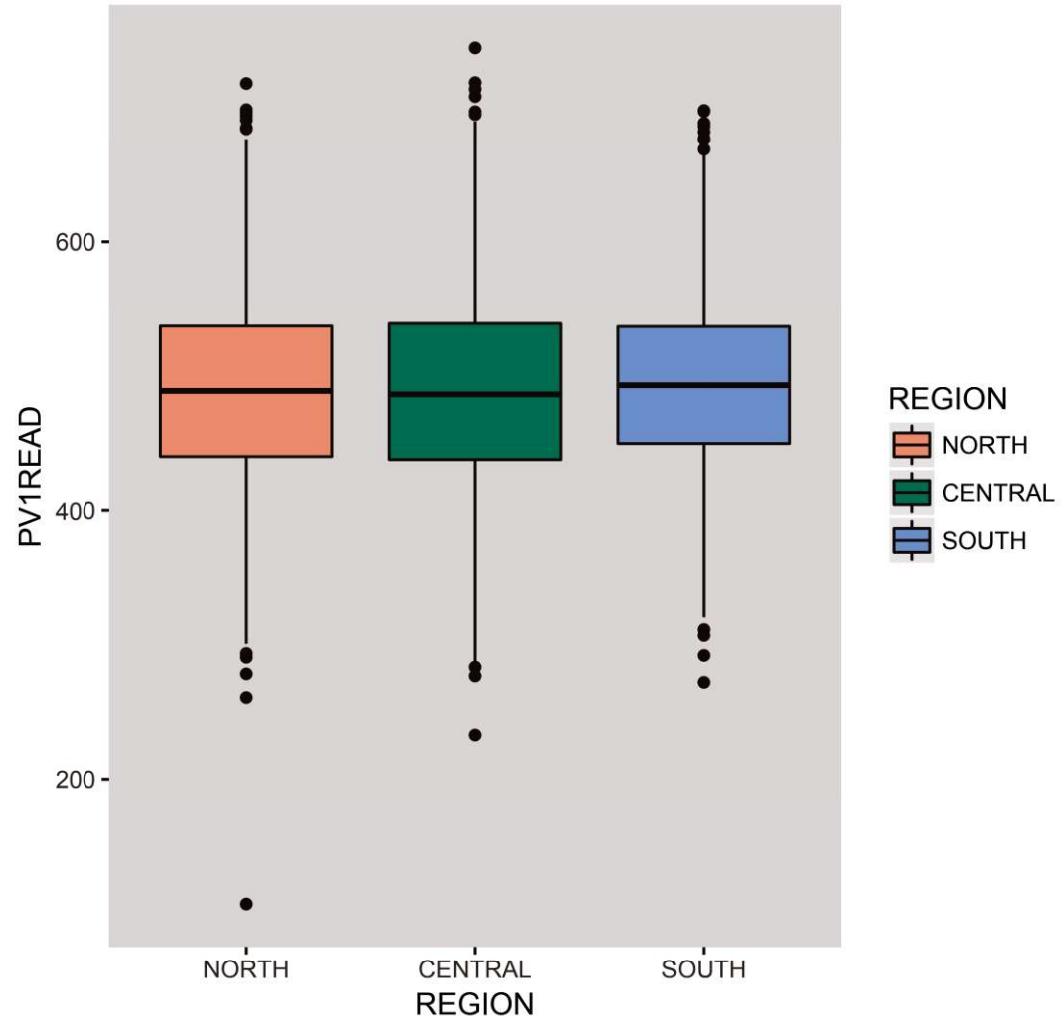
```
# vẽ biểu đồ theo Hisced
```

```
p = ggplot(pisa, aes(x=Hisced,  
y=PV1READ, color=Hisced))  
p = p + geom_boxplot()  
p
```



Biểu đồ hộp dùng ggplot2: đơn giản

```
# vẽ biểu đồ theo Hisced  
p = ggplot(pisa, aes(x=REGION,  
y=PV1READ, fill=REGION))  
p = p + geom_boxplot()  
p
```



Đến phiên bạn ...

```
install.packages("ggthemes")
library(ggthemes)
fill = "#4271AE"
line = "#1F3552"

p = ggplot(pisa, aes(x=Hisced, y=PV1READ, color=Hisced))
p = p + geom_boxplot(fill=fill, colour=line)
p + theme_economist()
```

Đến phiên bạn ... (điểm theo giới tính và Hisced)

```
p = ggplot(pisa, aes(x=Hisced, y=PV1READ, fill=Hisced))  
p = p + geom_boxplot(aes(fill=Hisced), alpha=1)  
p = p + geom_jitter(aes(color=Hisced, outlier.colour="red") ,  
size=1.5, alpha=.2)  
  
p = p + theme(axis.title.x = element_text(color="blue",  
size=10, face="bold"), axis.text.x = element_text(angle=45,  
vjust=0.5, size=10)) + xlab("HISCED") + ylab("Điểm Reading")  
  
p + theme_economist()
```

Đến phiên bạn ... (điểm theo giới tính và Hisced)

```
p = ggplot(pisa, aes(x=Hisced, y=PV1READ, fill=Hisced))  
p = p + geom_boxplot(aes(fill=Hisced), alpha=1)  
p = p + geom_jitter(aes(color=Hisced, outlier.colour="red") ,  
size=1.5, alpha=.2)  
  
p = p + theme(axis.title.x = element_text(color="blue",  
size=10, face="bold"), axis.text.x = element_text(angle=45,  
vjust=0.5, size=10)) + xlab("HISCED") + ylab("Điểm Reading")  
  
p + facet_wrap(~REGION) + theme_economist()
```

scatterplot (biểu đồ tương quan)

Liên quan giữa PV1SCIE và WEALTH

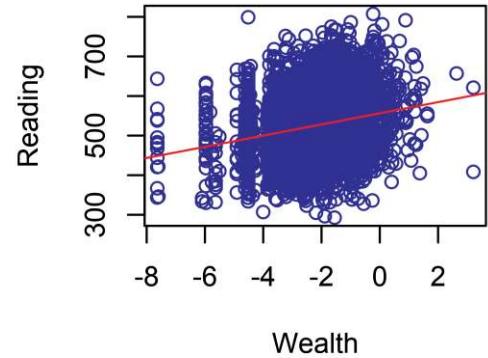
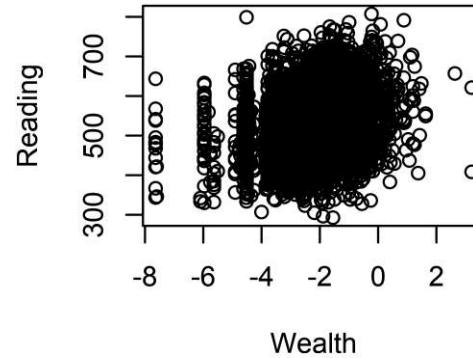
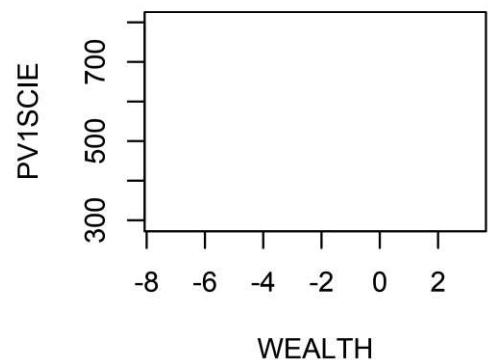
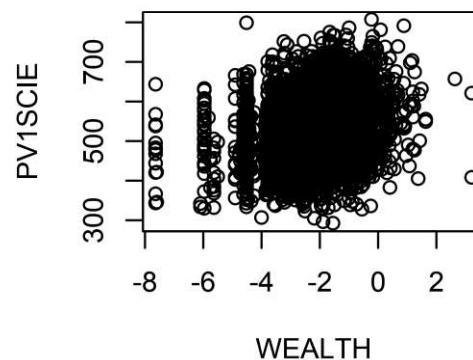
```
par(mfrow=c(2, 2))
```

```
plot(WEALTH, PV1SCIE)
```

```
plot(PV1SCIE ~ WEALTH, pch=16)
```

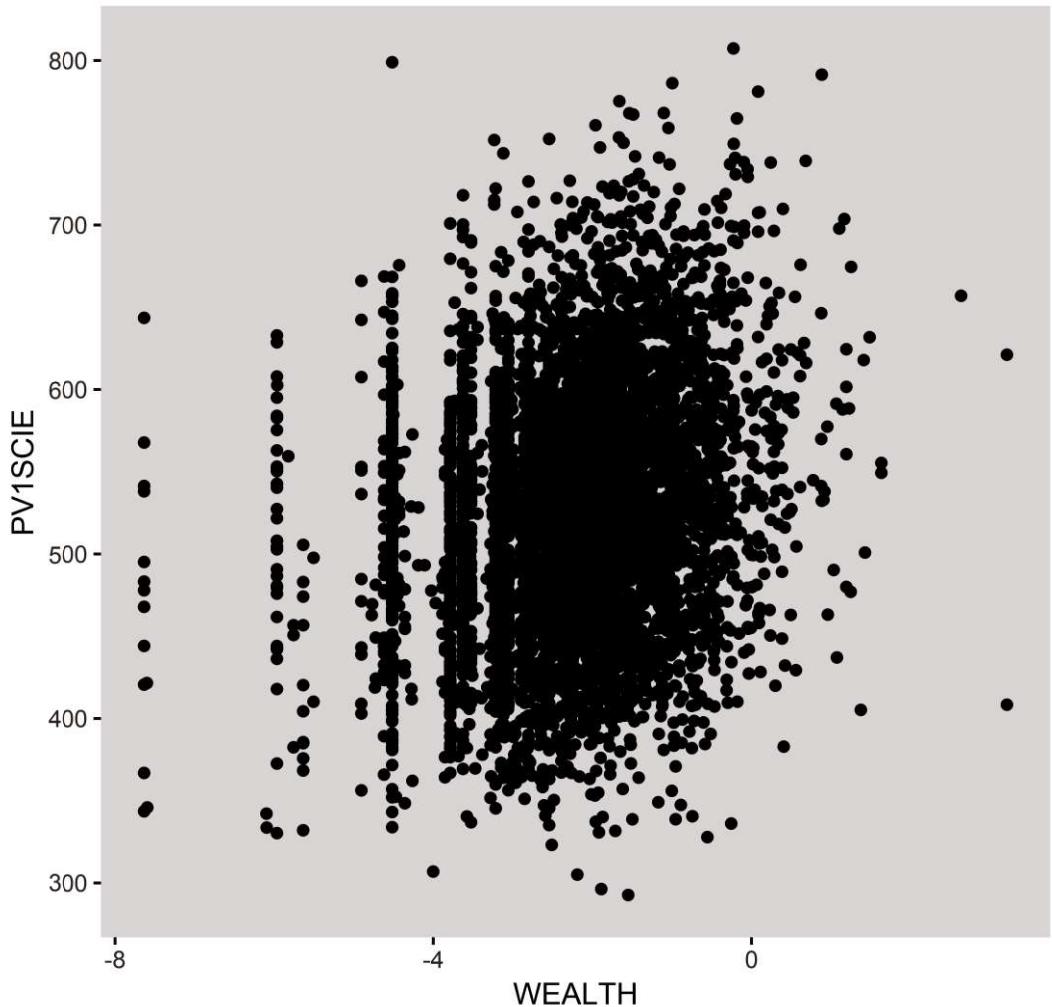
```
plot(PV1SCIE ~ WEALTH,  
xlab="Wealth", ylab="Reading")
```

```
plot(PV1SCIE ~ WEALTH,  
xlab="Wealth", ylab="Reading",  
col="blue")  
abline(lm(PV1SCIE ~ WEALTH),  
col="red")
```



Biểu đồ tương quan với ggplot2

```
library(ggplot2);  
library(gridExtra)  
  
p = ggplot(pisa, aes(x=WEALTH,  
y=PVI_SCIIE))  
p = p + geom_point()  
p
```



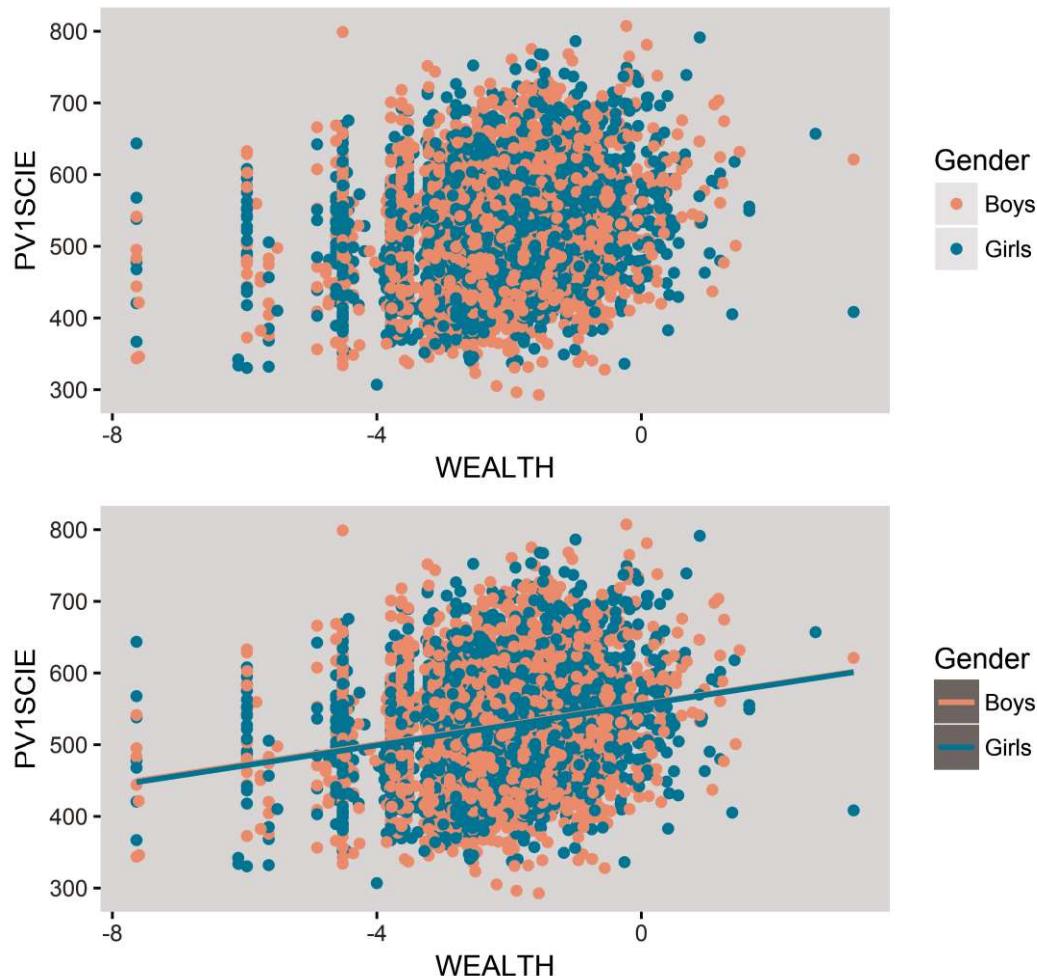
Biểu đồ tương quan với ggplot2: tô màu theo Gender

```
p = ggplot(pisa, aes(x=WEALTH,  
y=PV1SCIE, color=Gender))
```

```
p1 = p + geom_point()
```

```
p2 = p + geom_point() +  
geom_smooth(method="lm")
```

```
grid.arrange(p1, p2, nrow=2)
```



Đến phiên bạn ... PV1READ theo giới tính và tuổi

```
p = ggplot(pisa, aes(x=AGE, y=PV1READ, color=Gender))  
p1 = p + geom_point()  
  
p2 = p + geom_point() + geom_smooth(method="lm")  
  
grid.arrange(p1, p2, nrow=2)
```

Tóm tắt

Phân tích dữ liệu bằng biểu đồ

- RẤT quan trọng!
- Cần có sẵn câu hỏi nghiên cứu
- Triển khai câu hỏi nghiên cứu
 - hàm R (hist, boxplot, plot, barplot, ggplot2)
 - tên biến số
- Loại biểu đồ:
 - mô tả tần số: barplot
 - phân bố: histogram
 - khác biệt: boxplot
 - tương quan: scatterplot
 - sống còn: survival plot
- Biểu đồ cần mang tính "professional" và thẩm mĩ

Đến phiên bạn ...

Dữ liệu PISA: Misced theo vùng

```
pisa = read.csv("~/Dropbox/_Conferences and Workshops/Hanoi  
2017/PISA VN 2015.csv")  
  
attach(pisa)  
  
# tạo biến factor  
pisa$Misced = as.factor(pisa$MISCED)  
pisa$Pared = as.factor(pisa$PARED)  
  
# vẽ biểu đồ  
  
p = ggplot(data=pisa, aes(x=Misced, fill=REGION))  
p = p + geom_bar(aes(y = ..count../sum(..count..)))  
p + xlab("MISCED")
```

Dữ liệu PISA: Misced theo vùng

```
# vẽ biểu đồ theo tuổi và giới  
tính stackbar
```

```
p = ggplot(data=pisa, aes(x=PARED,  
y=AREA, fill=Gender))
```

```
p = p + geom_bar(aes(y =  
(..count..)/sum(..count..)))
```

```
p + xlab("PARED")
```

```
# vẽ biểu đồ theo tuổi và giới  
tính với dodge
```

```
p = ggplot(data=pisa,  
aes(x=PARED, fill=AREA))
```

```
p = p +  
geom_bar(position="dodge", aes(y  
= (..count..)/sum(..count..)))
```

```
p + xlab("PARED")
```

Biểu đồ hộp: điểm toán theo PARED

```
# install.packages("ggthemes")  
library(ggthemes)  
fill = "#4271AE"  
line = "#1F3552"  
  
p = ggplot(pisa, aes(x=as.factor(PARED), y=PV1MATH))  
p = p + geom_boxplot(fill=fill, colour=line)  
p + theme_economist()
```

Biểu đồ hộp: điểm toán theo PARED và giới tính

```
p = ggplot(pisa, aes(x=REGION, y=PV1SCIE, fill=Gender))
p = p + geom_boxplot(aes(fill=Gender), alpha=1)
p = p + geom_jitter(aes(color=Gender, outlier.colour="red"), size=1.5,
alpha=.2)

p = p + theme(axis.title.x = element_text(color="blue", size=10,
face="bold"), axis.text.x  = element_text(angle=45, vjust=0.5,
size=10))

p + theme_economist()
```

```
p = ggplot(pisa, aes(x=AREA, y=PV1READ, fill=Gender))
p = p + geom_boxplot(aes(fill=Gender), alpha=1)
p = p + geom_jitter(aes(color=Gender, outlier.colour="red"), size=1.5,
alpha=.2)

p = p + theme(axis.title.x = element_text(color="blue", size=10,
face="bold"), axis.text.x  = element_text(angle=45, vjust=0.5,
size=10))

p + facet_wrap(~Gender) + theme_economist()
```

Điểm môn PV1SCIE theo WEALTH và giới tính

```
library(gridExtra)

p = ggplot(pisa, aes(x=WEALTH, y=PV1SCIE, color=Gender) )
p1 = p + geom_point()

p2 = p + geom_point() + geom_smooth(method="lm")

grid.arrange(p1, p2, nrow=2)
```

Đến phiên bạn ... tỉ trọng mỡ theo bmi và giới tính

```
# ob = read.csv("~/Dropbox/_Conferences and Workshops/Da Nang 2017/  
Datasets/obesity data.csv")  
ob = read.csv(file.choose())  
  
p = ggplot(ob, aes(x=bmi, y=pcfat, fill=gender, col=gender))  
p = p + geom_point() + geom_smooth(method="lm", formula=y~x+I(x^2))  
p + xlab("Body mass index") + ylab("Percent body fat")
```

Đến phiên bạn ... phân tích phiếu tín nhiệm

```
# tn = read.csv("~/Dropbox/_Conferences and Workshops/Da Nang 2017/  
Datasets/Lay phieu tin nhiem.csv")  
tn = read.csv(file.choose())  
  
p = ggplot(tn, aes(x=Year2013, y=Year2014, col=Group))  
p = p + geom_point(shape=16, size=2)  
p = p + geom_text(aes(label=Name), size=5)  
p = p + geom_abline(slope=1, size=0.5, lty=2)  
p + theme_bw()
```