Chapter 12: ggplot2

Exercise 1: Chol - ggplot2

- Cho dữ liệu chol.txt.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ histogram cho cột AGE của dữ liệu
- Vẽ scatter plot biểu diễn mối quan hệ của HEIGHT vs WEIGHT
- Vẽ scatter plot biểu diễn mối quan hệ của HEIGHT vs WEIGHT, có kèm theo Histogram / Boxplot
- Vẽ piechart biểu diễn BLOOD

Exercise 2: Housing Prices

- Cho dữ liệu landdata_states.csv.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ histogram cho cột Home.Value của dữ liệu
- Hãy lọc dữ liệu theo Date == 2001.25, sau đó vẽ scatter plot biểu diễn Land.Value vs Structure.Cost.
- Vẽ lại biểu đồ trên với Land.Value được chuẩn hóa bằng log. Gắn thêm state cho từng điểm dữ liệu.

Exercise 3: EconomistData

- Cho dữ liệu EconomistData.csv.
- Đọc và hiển thị head của dữ liệu
- In thống kê chung về dữ liệu
- In thông tin của dữ liệu
- Vẽ scatter plot biểu diễn mối quan hệ của CPI vs HDI, điểm được tô màu theo Region, độ lớn của điểm theo HDI.Rank
- Vẽ scatter plot biểu diễn mối quan hệ của CPI vs HDI có regression line

---

Gợi ý:

Exercise 1: Chol - ggplot2

In [1]:
```
library(ggplot2)
```

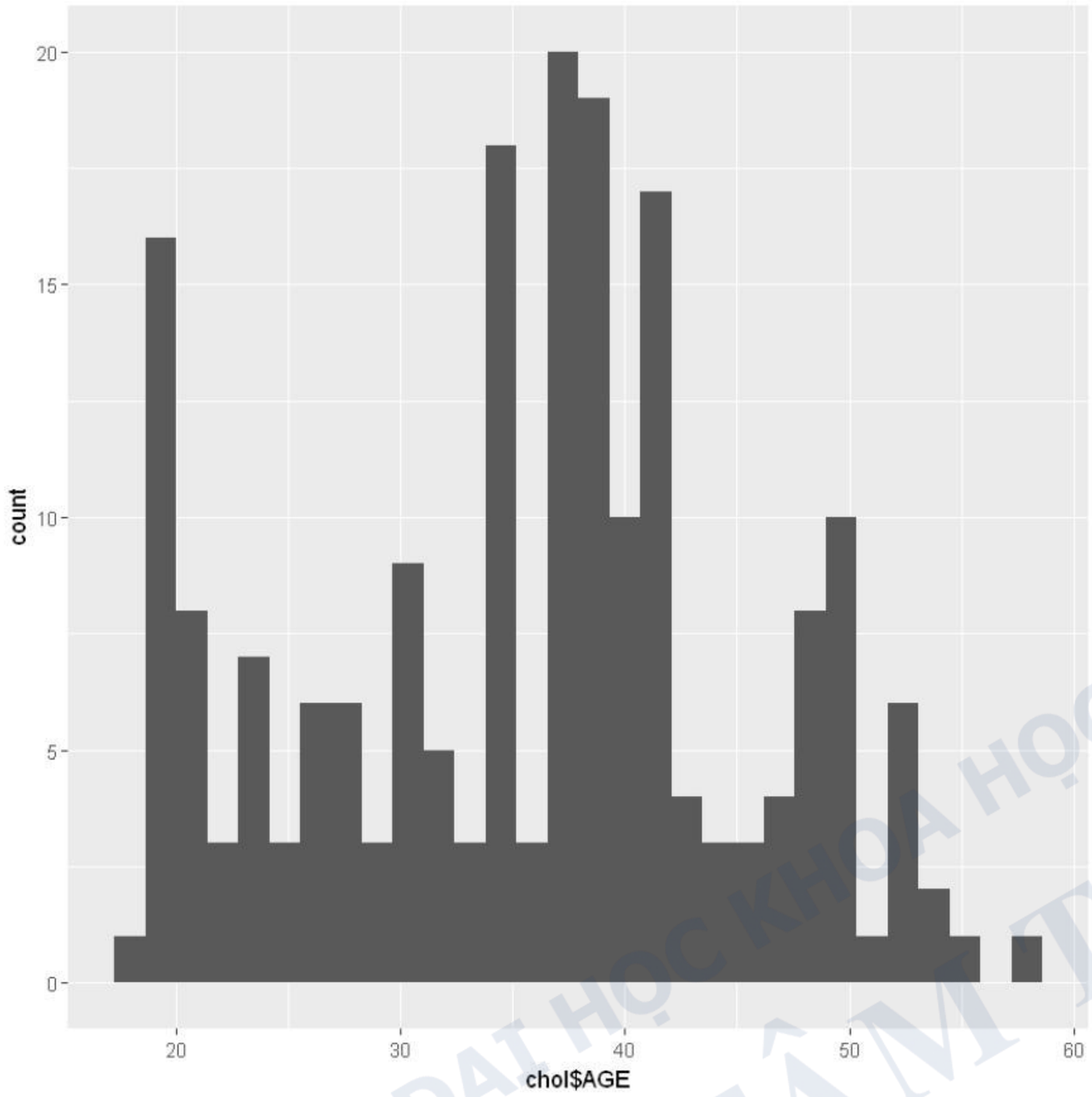
In [2]:
```
# Load in `chol` data
chol <- read.table("chol.txt", header = TRUE)

# Inspect first rows of `chol` with `head()`
head(chol)

# Summary with `summary()`
summary(chol)

# Structure of `chol` with `str()`
str(chol)
```

AGE	HEIGHT	WEIGHT	CHOL	SMOKE	BLOOD	MORT
20	176	77	195	nonsmo	b	alive
53	167	56	250	sigare	o	dead
44	170	80	304	sigare	a	dead
37	173	89	178	nonsmo	o	alive
26	170	71	206	sigare	o	alive
41	165	62	284	sigare	o	alive

```
      AGE            HEIGHT          WEIGHT           CHOL          SMOKE    
 Min.   :18.00   Min.   :156.0   Min.   : 53.00   Min.   :107.0   nonsmo: 49  
 1st Qu.:28.75   1st Qu.:168.0   1st Qu.: 68.75   1st Qu.:204.0   pipe  : 42  
 Median :37.00   Median :172.0   Median : 75.00   Median :232.0   sigare:109  
 Mean   :35.72   Mean   :172.3   Mean   : 75.89   Mean   :233.6               
 3rd Qu.:42.00   3rd Qu.:176.0   3rd Qu.: 82.00   3rd Qu.:259.0               
 Max.   :58.00   Max.   :191.0   Max.   :110.00   Max.   :455.0               
 BLOOD       MORT      
 a :82   alive:176  
 ab: 5   dead : 24  
 b :22              
 o :91              
```

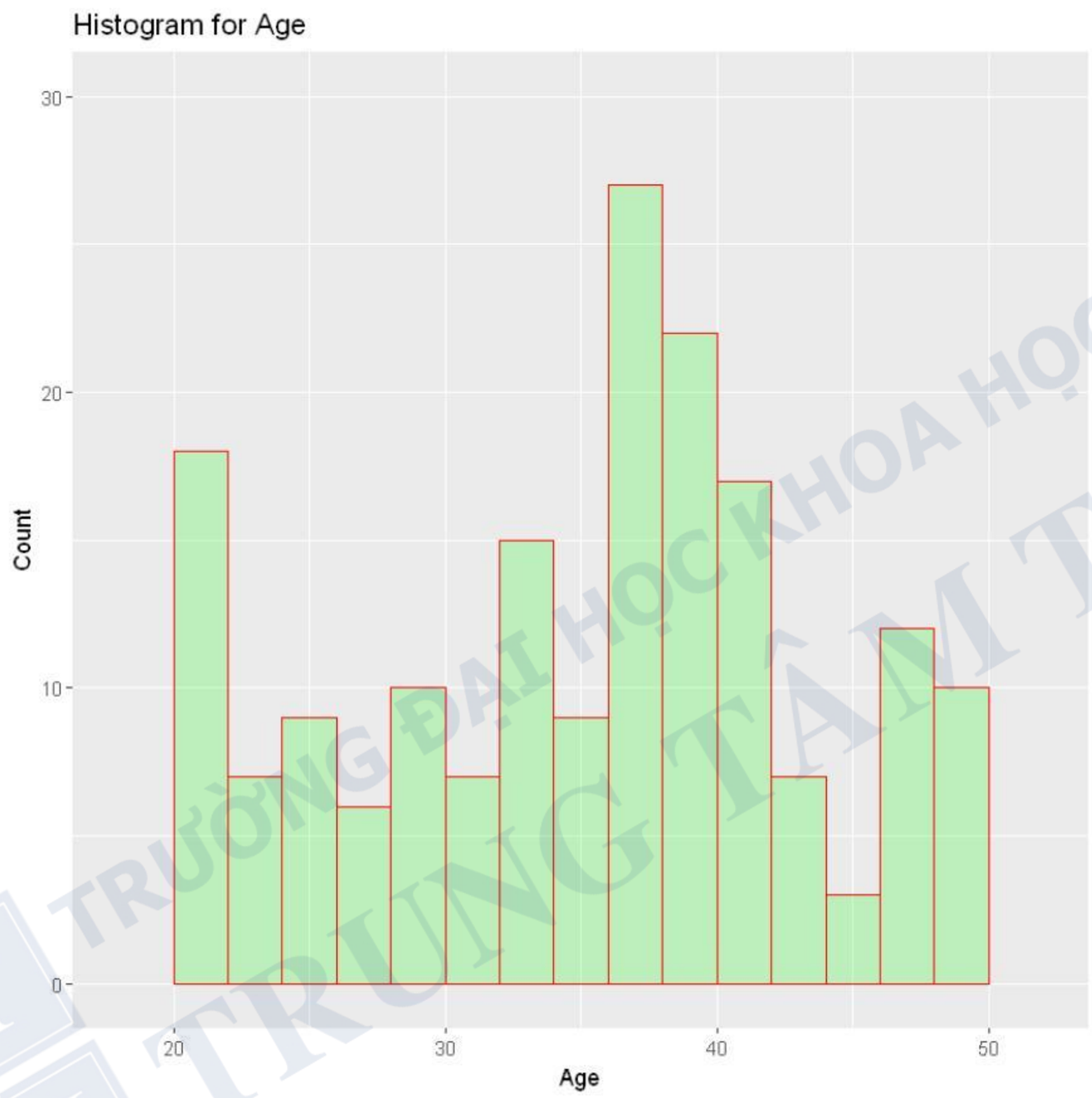
```
'data.frame':	200 obs. of  7 variables:
 $ AGE   : int  20 53 44 37 26 41 39 28 33 39 ...
 $ HEIGHT: int  176 167 170 173 170 165 174 171 180 166 ...
 $ WEIGHT: int  77 56 80 89 71 62 75 68 100 74 ...
 $ CHOL  : int  195 250 304 178 206 284 232 152 209 150 ...
 $ SMOKE : Factor w/ 3 levels "nonsmo","pipe",..: 1 3 3 1 3 3 3 2 3 3 ...
 $ BLOOD : Factor w/ 4 levels "a","ab","b","o": 3 4 1 4 4 4 4 1 1 1 ...
 $ MORT  : Factor w/ 2 levels "alive","dead": 1 2 2 1 1 1 1 1 1 1 ...
```

In [8]:
```
ggplot(data=chol, aes(chol$AGE)) +
  geom_histogram(bins=30)
```

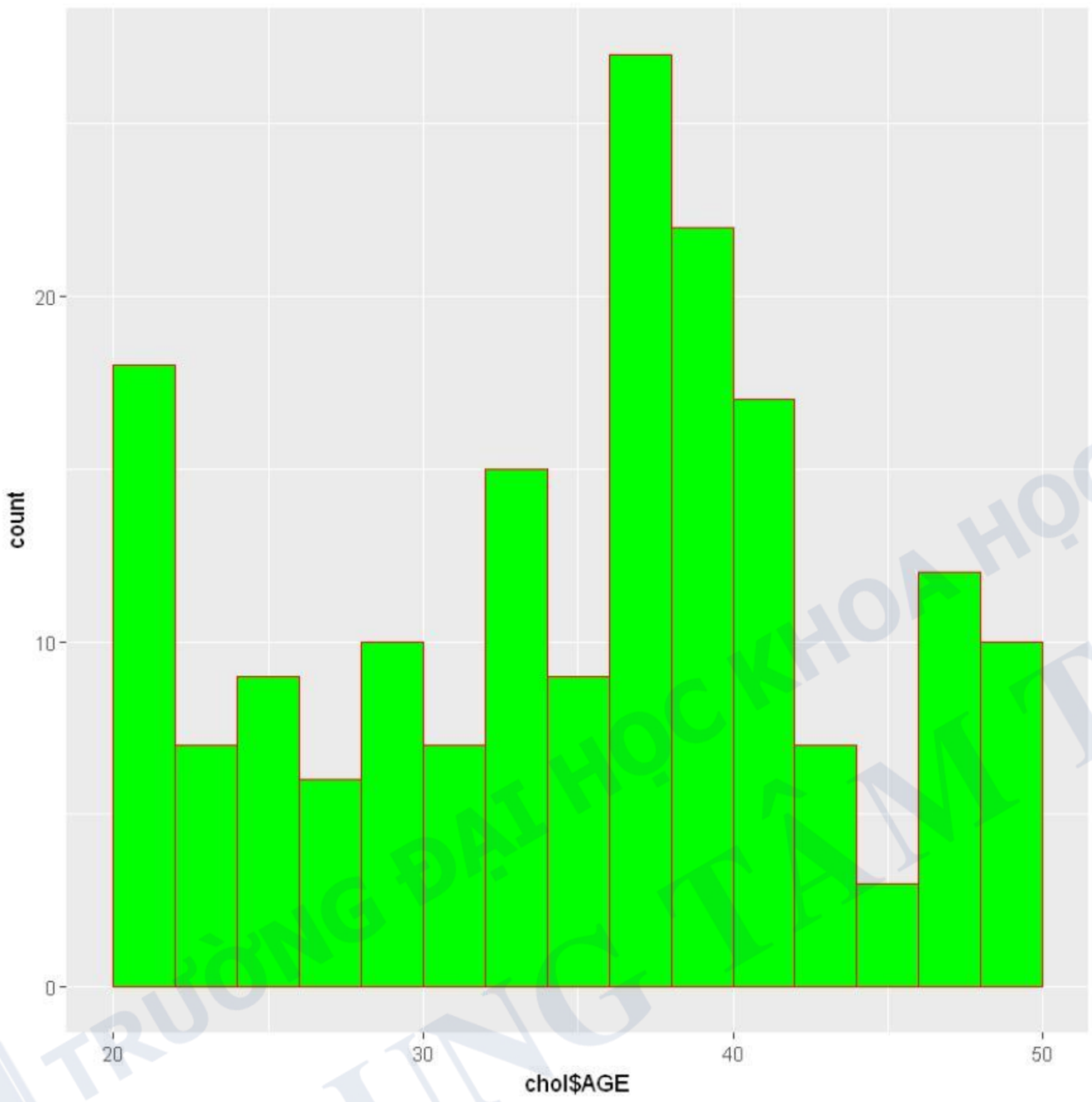


In [11]:
```
ggplot(data=chol, aes(x=chol$AGE)) +
  geom_histogram(breaks=seq(20, 50, by=2),
                 col="red",
                 fill="green",
                 alpha = .2) +
  labs(title="Histogram for Age", x="Age", y="Count") +
  xlim(c(18,52)) +
  ylim(c(0,30))
```

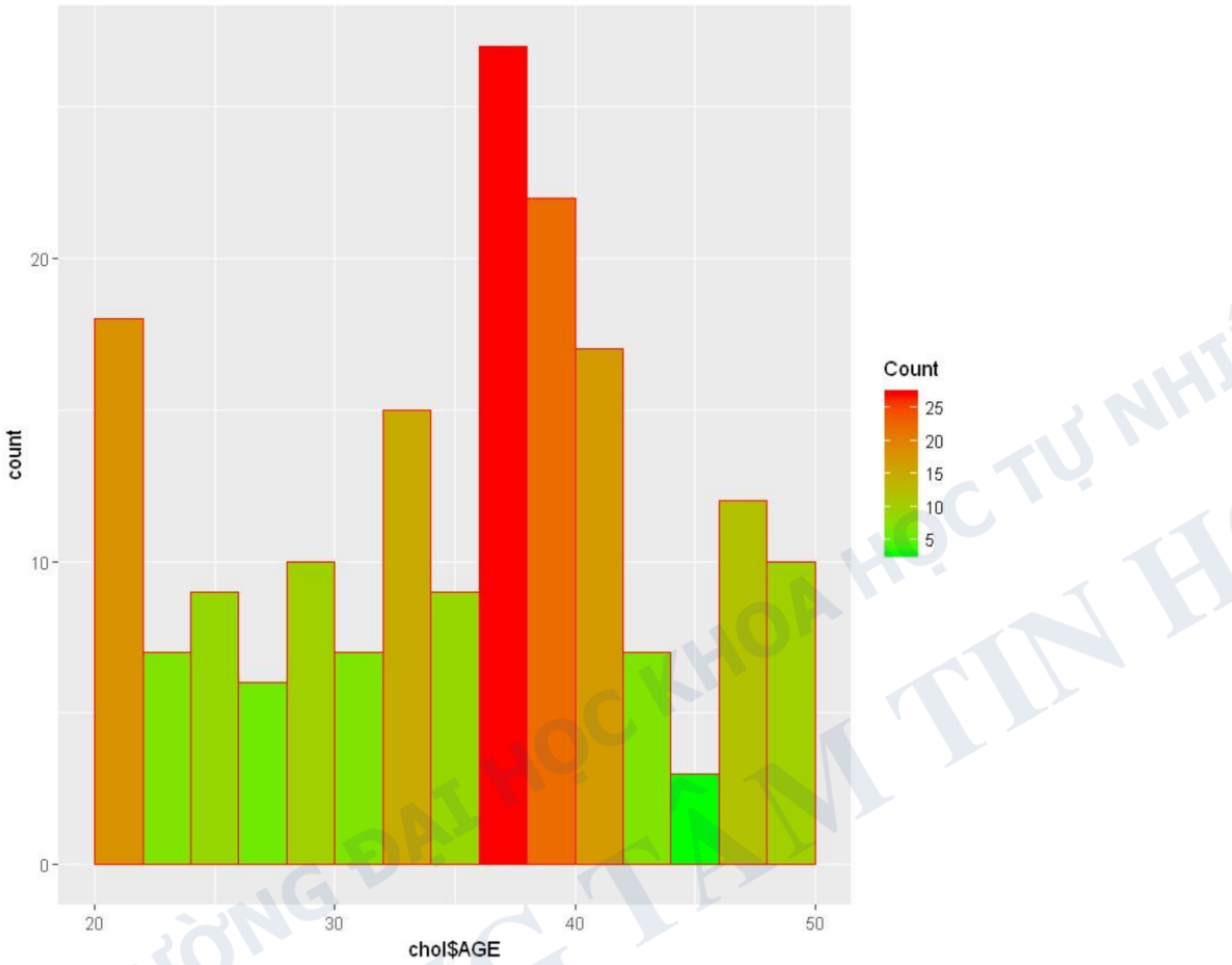
```
Warning message:
"Removed 6 rows containing non-finite values (stat_bin)."
```



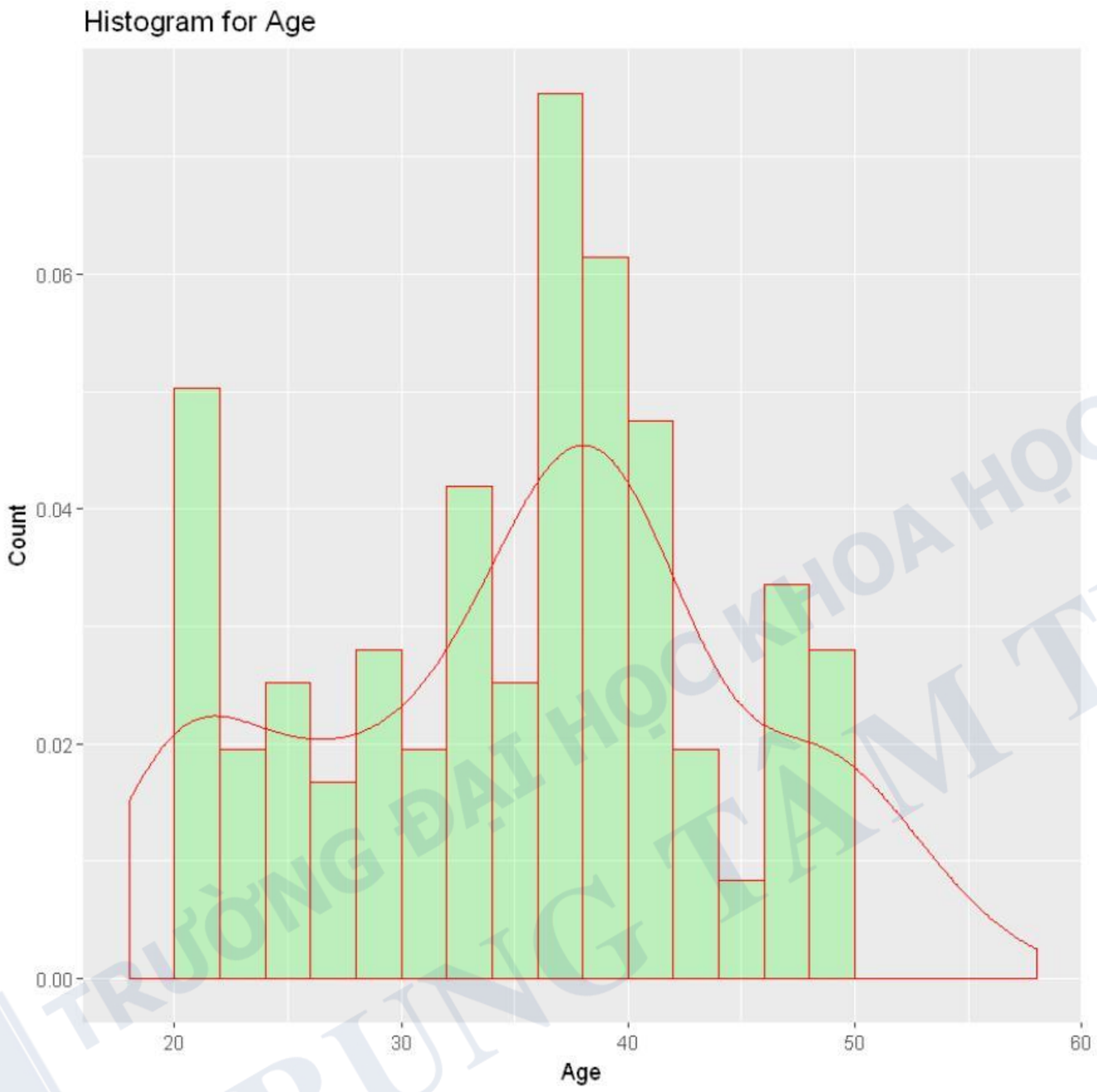
In [12]:
```
ggplot(data=chol, aes(chol$AGE)) +
  geom_histogram(breaks=seq(20, 50, by=2),
                 col="red",
                 fill="green")
```



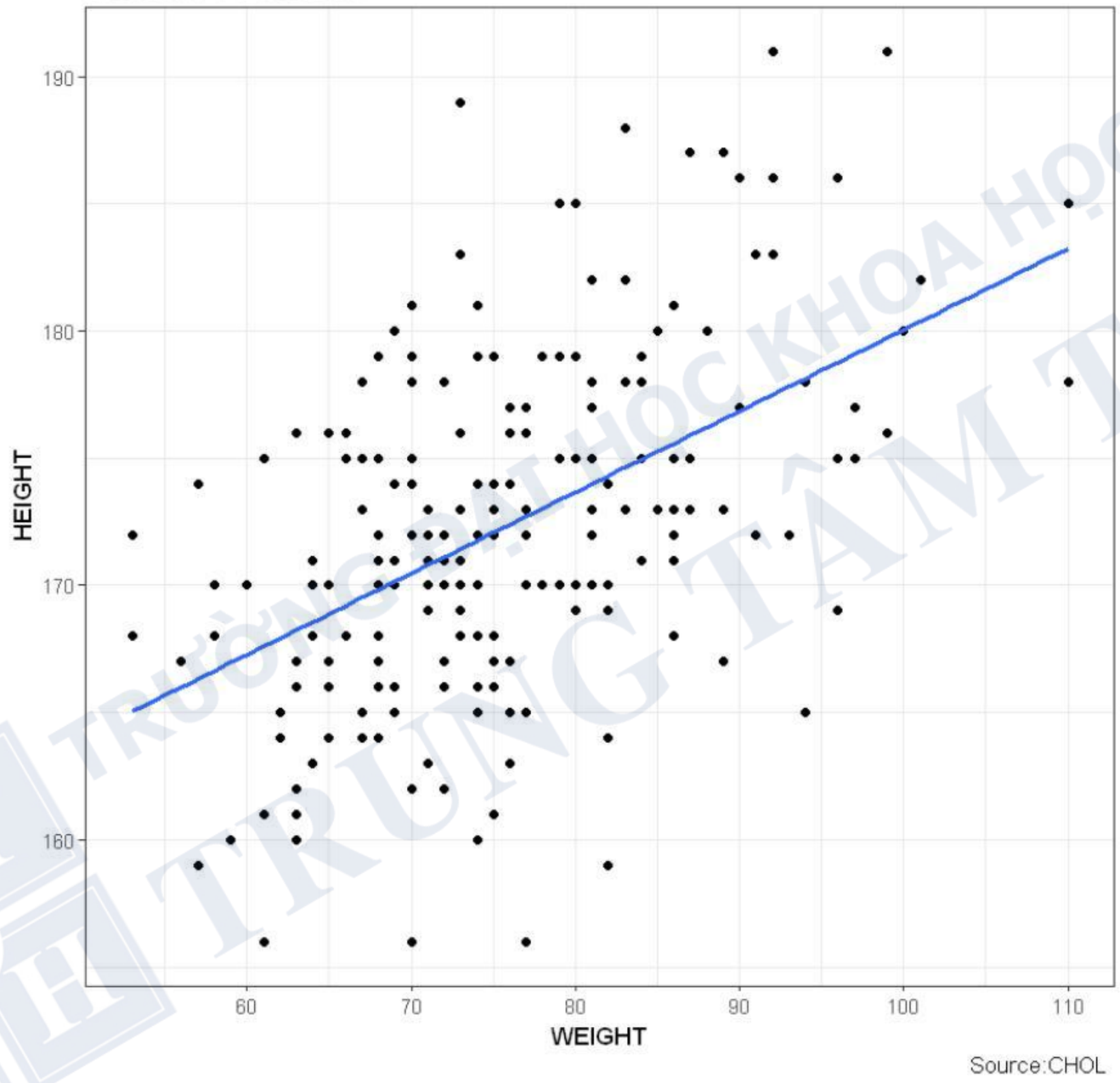
In [13]:
```
ggplot(data=chol, aes(chol$AGE)) +
  geom_histogram(breaks=seq(20, 50, by=2),
                 col="red",
                 aes(fill=..count..)) +
  scale_fill_gradient("Count", low="green", high="red")
```



In [14]:
```
ggplot(data=chol, aes(chol$AGE)) +
  geom_histogram(aes(y =..density..),
                 breaks=seq(20, 50, by = 2),
                 col="red",
                 fill="green",
                 alpha=.2) +
  geom_density(col=2) +
  labs(title="Histogram for Age", x="Age", y="Count")
```

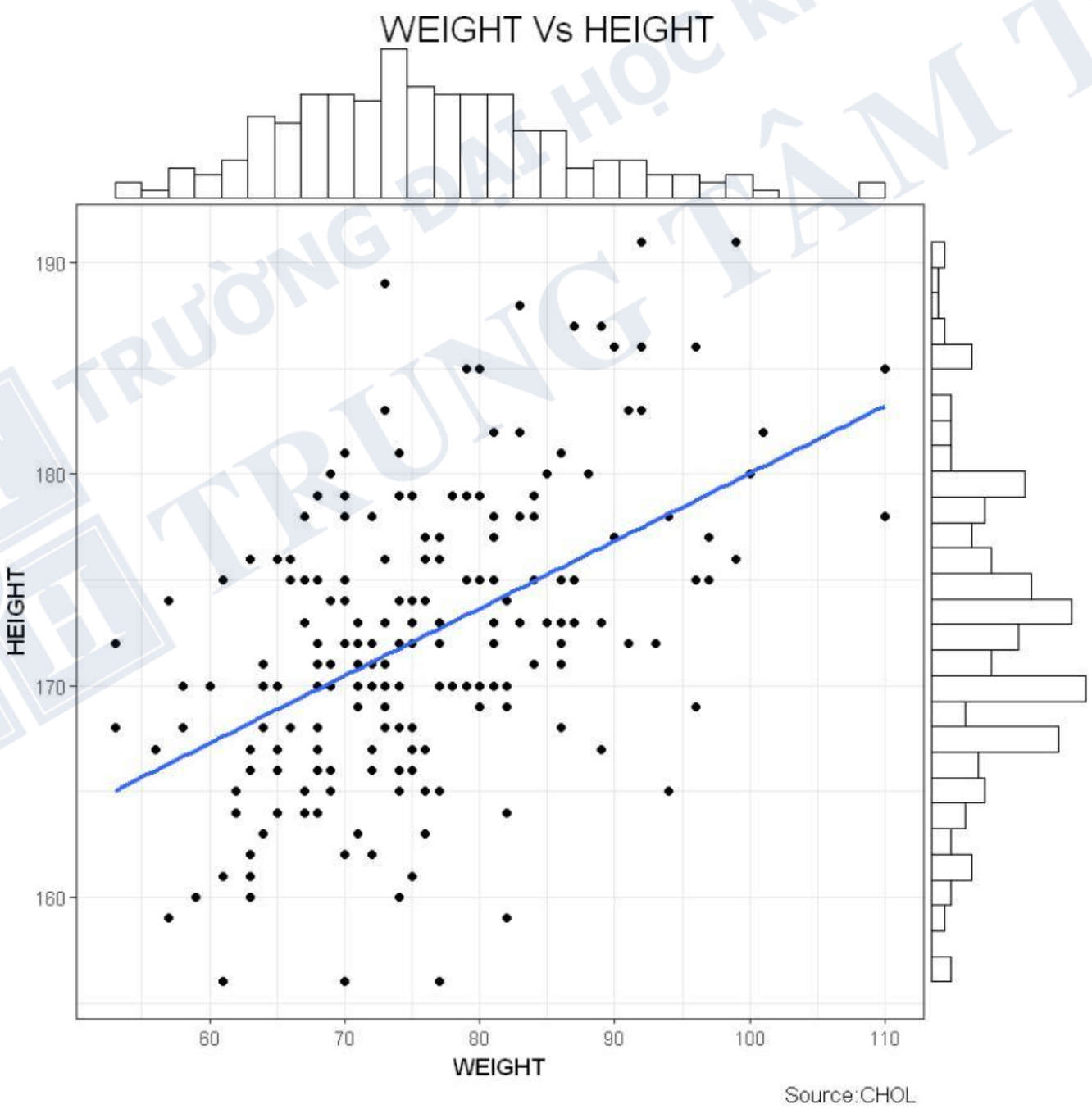


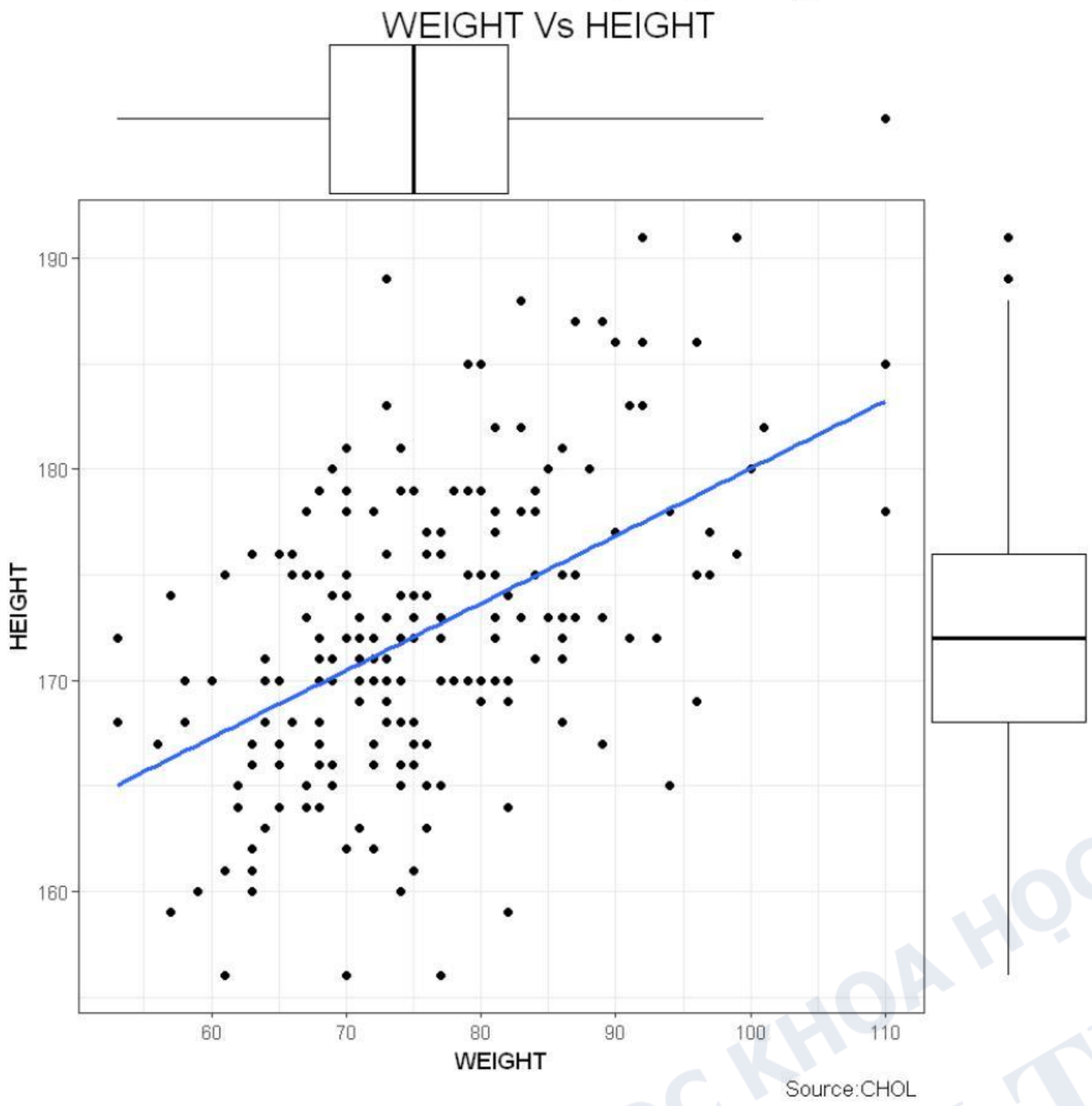
In [21]:
```
options(scipen=999)  # turn-off scientific notation like 1e+48
library(ggplot2)
theme_set(theme_bw())  # pre-set the bw theme.

# Scatterplot
gg <- ggplot(data = chol, aes(x=WEIGHT, y=HEIGHT)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  labs(y="HEIGHT",
       x="WEIGHT",
       title="WEIGHT Vs HEIGHT",
       caption = "Source:CHOL ")

plot(gg)
```



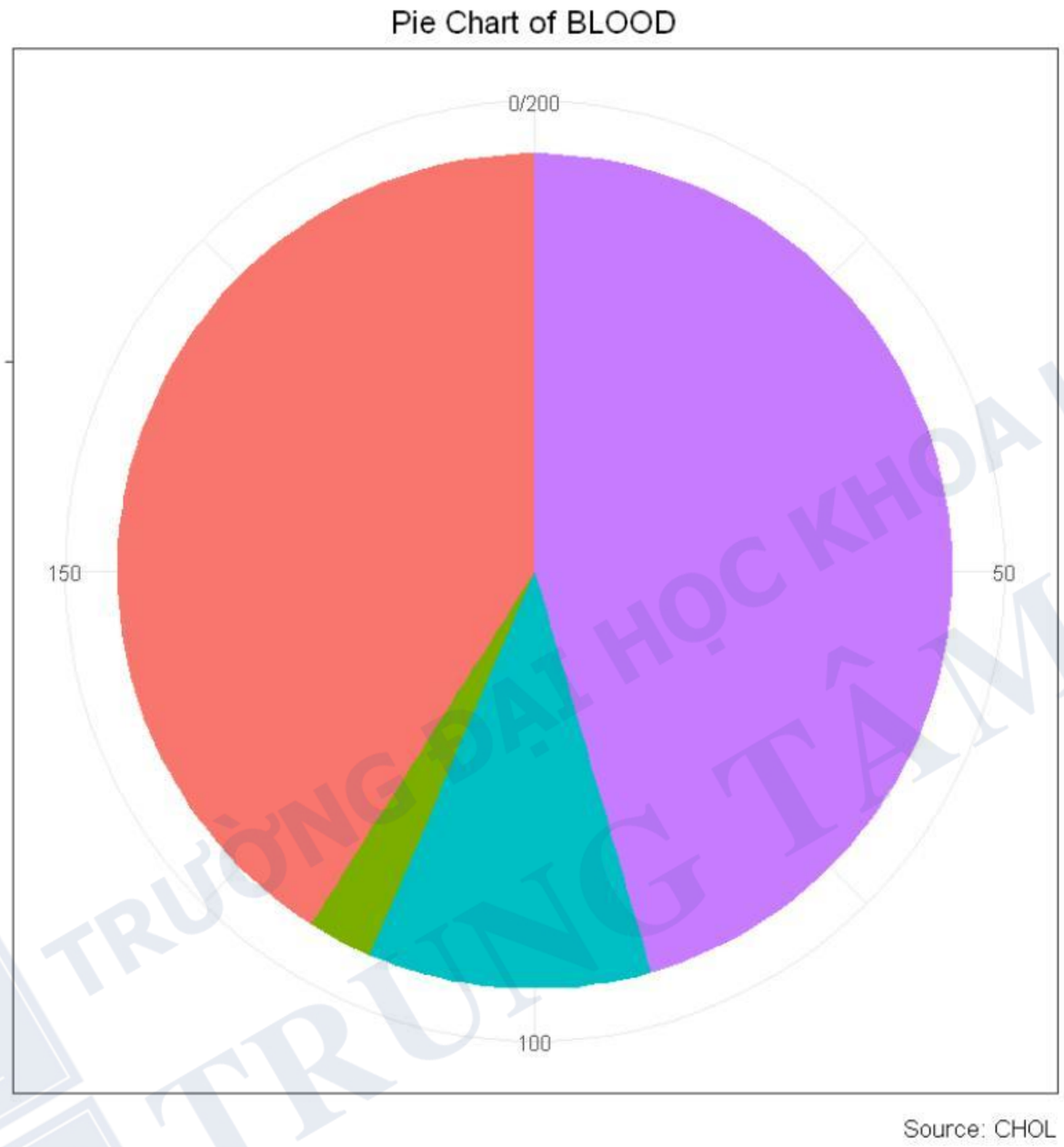
In [23]:
```
# Load package and data

library(ggExtra)
theme_set(theme_bw())  # pre-set the bw theme.

# Scatterplot
gg <- ggplot(data = chol, aes(x=WEIGHT, y=HEIGHT)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  labs(y="HEIGHT",
       x="WEIGHT",
       title="WEIGHT Vs HEIGHT",
       caption = "Source:CHOL ")

ggMarginal(gg, type = "histogram", fill="transparent", bins = 30)
ggMarginal(gg, type = "boxplot", fill="transparent")
```

```
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
Warning message:
"Continuous x aesthetic -- did you forget aes(group=...)?"
```





In [26]:
```
pie <- ggplot(chol, aes(x = "", fill = factor(BLOOD))) +
  geom_bar(width = 1) +
  theme(axis.line = element_blank(),
        plot.title = element_text(hjust=0.5)) +
  labs(fill="class",
       x=NULL,
       y=NULL,
       title="Pie Chart of BLOOD",
       caption="Source: CHOL")

pie + coord_polar(theta = "y", start=0)
```



Exercise 2: Housing Prices

In [30]:
```
housing = read.csv("landdata_states.csv")
```

In [32]:
```
head(housing)
```

State	region	Date	Home.Value	Structure.Cost	Land.Value	Land.Share..Pct.	Home.Price.Index
AK	West	2010.25	224952	160599	64352	28.6	1.481
AK	West	2010.50	225511	160252	65259	28.9	1.484
AK	West	2009.75	225820	163791	62029	27.5	1.486
AK	West	2010.00	224994	161787	63207	28.1	1.481
AK	West	2008.00	234590	155400	79190	33.8	1.544
AK	West	2008.25	233714	157458	76256	32.6	1.538

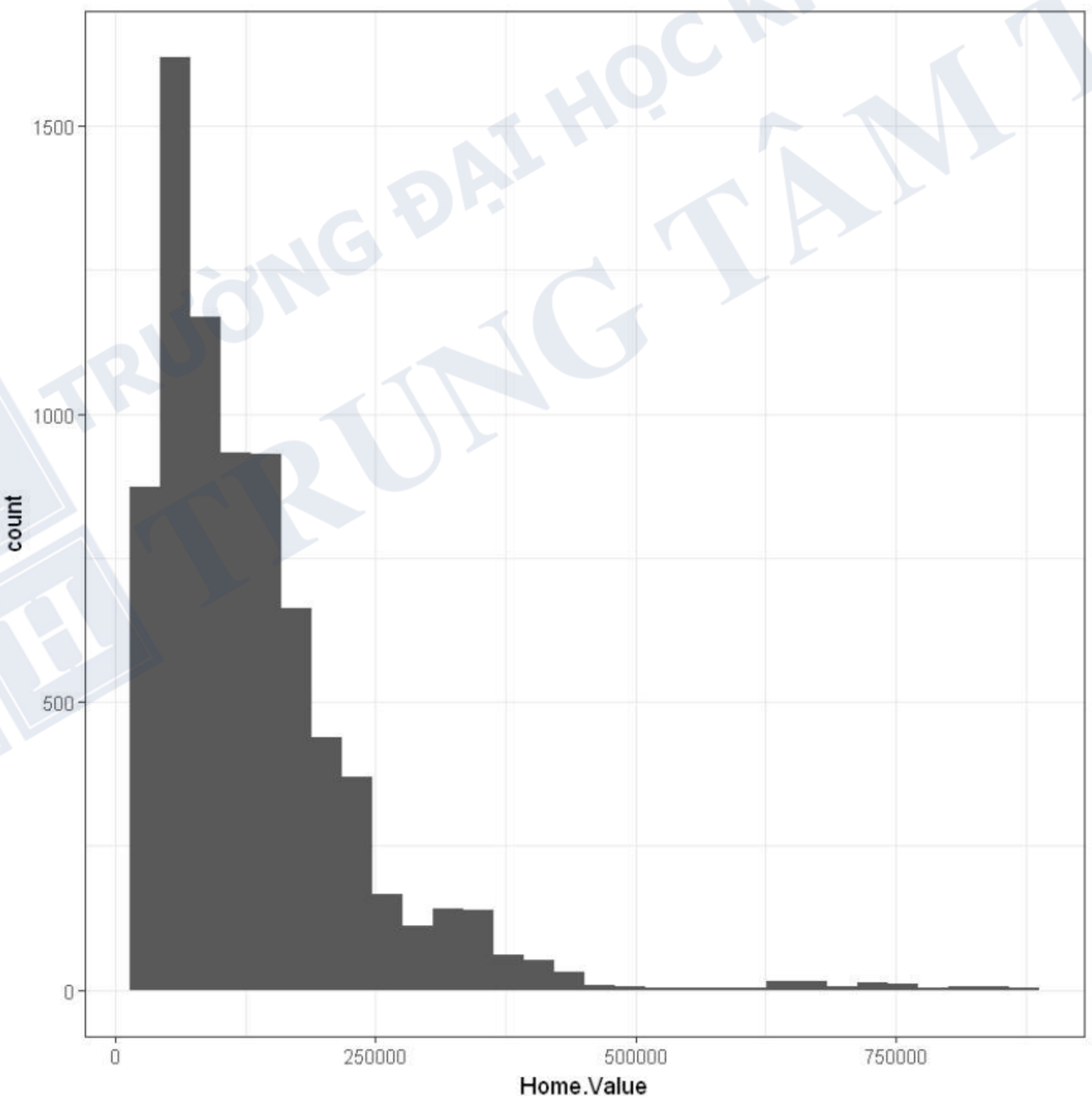
In [34]:
```
summary(housing)
```

```
       State             region          Date        Home.Value     Structure.Cost 
 AK     : 153   Midwest:1836   Min.   :1975   Min.   : 18763   Min.   : 17825  
 AL     : 153   N. East:1377   1st Qu.:1985   1st Qu.: 62235   1st Qu.: 53777  
 AR     : 153   South  :2448   Median :1994   Median :108724   Median : 88352  
 AZ     : 153   West   :1989   Mean   :1994   Mean   :135313   Mean   : 99534  
 CA     : 153   NA's   : 153   3rd Qu.:2004   3rd Qu.:172031   3rd Qu.:134871  
 CO     : 153                  Max.   :2013   Max.   :862885   Max.   :325595  
 (Other):6885                                                                  
  Land.Value     Land.Share..Pct. Home.Price.Index Land.Price.Index
 Min.   :   938   Min.   : 5.00   Min.   :0.1350   Min.   : 0.0000  
 1st Qu.:  4178   1st Qu.: 5.00   1st Qu.:0.4550   1st Qu.: 0.0020  
 Median :  9478   Median :10.40   Median :0.7830   Median : 0.2520  
 Mean   : 35779   Mean   :18.17   Mean   :0.8695   Mean   : 0.9912  
 3rd Qu.: 38631   3rd Qu.:26.30   3rd Qu.:1.2075   3rd Qu.: 1.1510  
 Max.   :594417   Max.   :81.70   Max.   :2.8930   Max.   :15.4340  

      Year          Qrtr      
 Min.   :1975   Min.   :1.00  
 1st Qu.:1984   1st Qu.:1.00  
 Median :1994   Median :2.00  
 Mean   :1994   Mean   :2.49  
 3rd Qu.:2003   3rd Qu.:3.00  
 Max.   :2013   Max.   :4.00  
```

In [33]:
```
str(housing)
```

```
'data.frame':   7803 obs. of  11 variables:
 $ State           : Factor w/ 51 levels "AK","AL","AR",..: 1 1 1 1 1 1 1 1 1 1
...
 $ region          : Factor w/ 4 levels "Midwest","N. East",..: 4 4 4 4 4 4 4 4
4 4 ...
 $ Date            : num  2010 2010 2010 2010 2008 ...
 $ Home.Value      : int  224952 225511 225820 224994 234590 233714 232999 2321
64 231039 229395 ...
 $ Structure.Cost  : int  160599 160252 163791 161787 155400 157458 160092 1627
04 164739 165424 ...
 $ Land.Value      : int  64352 65259 62029 63207 79190 76256 72906 69460 66299
63971 ...
 $ Land.Share..Pct.: num  28.6 28.9 27.5 28.1 33.8 32.6 31.3 29.9 28.7 27.9 ...
 $ Home.Price.Index: num  1.48 1.48 1.49 1.48 1.54 ...
 $ Land.Price.Index: num  1.55 1.58 1.49 1.52 1.88 ...
 $ Year            : int  2010 2010 2009 2009 2007 2008 2008 2008 2008 2009 ...
 $ Qrtr            : int  1 2 3 4 4 1 2 3 4 1 ...
```

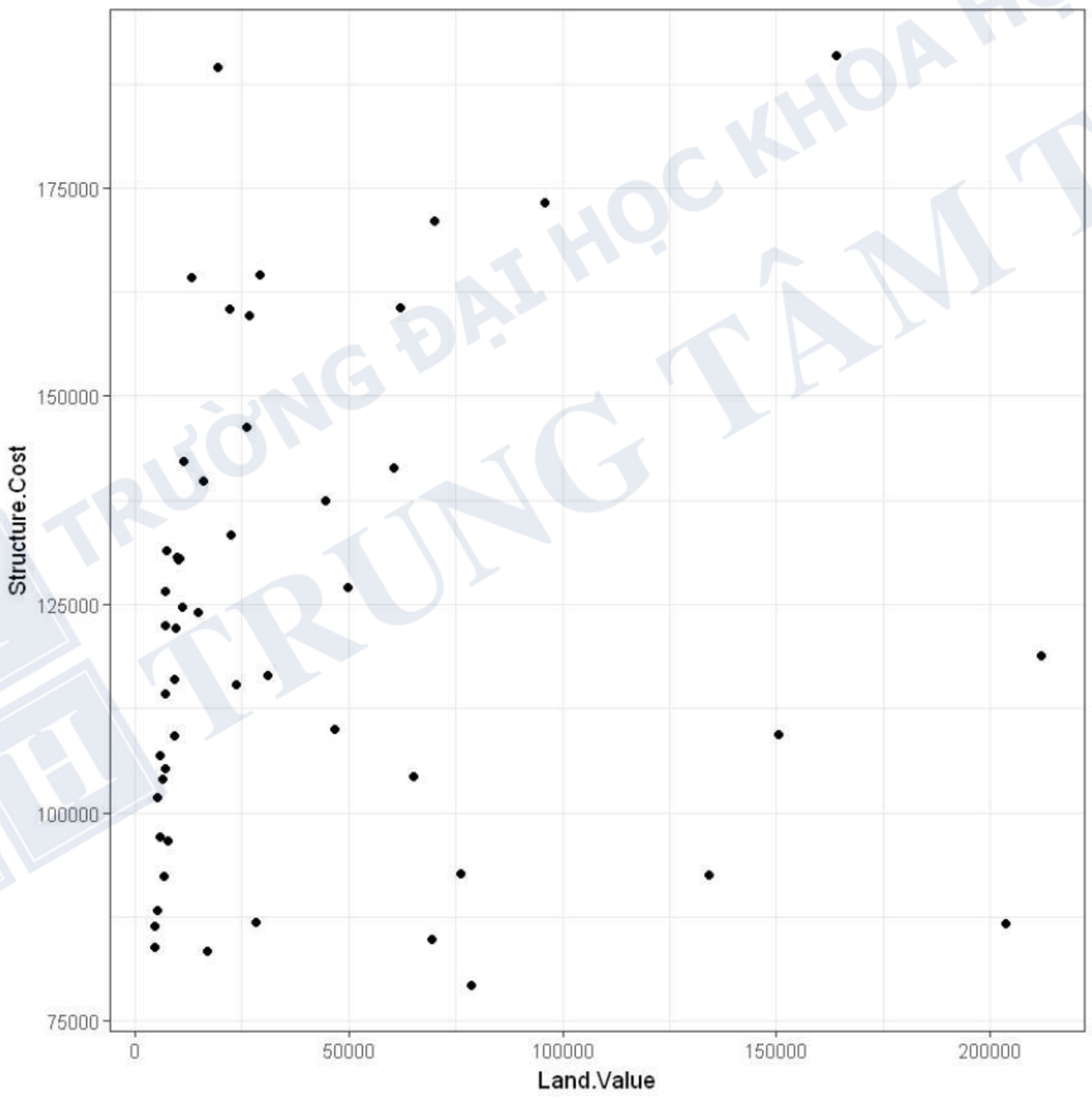
In [37]:
```
ggplot(housing, aes(x = Home.Value)) +
  geom_histogram(bins = 30)
```



In [40]:
```
library(tidyverse)
```

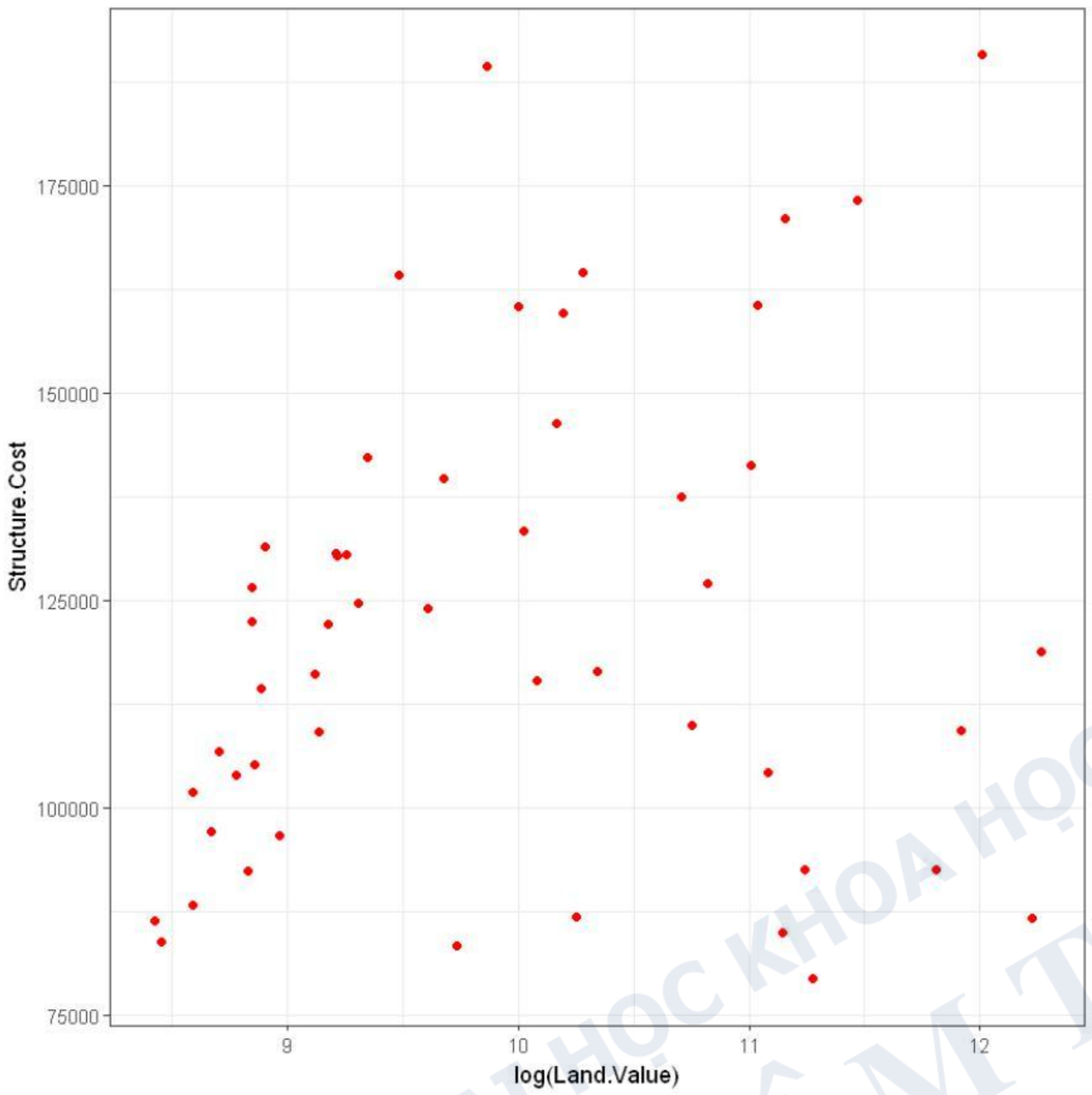
```
Loading tidyverse: tibble
Loading tidyverse: tidyr
Loading tidyverse: readr
Loading tidyverse: purrr
Loading tidyverse: dplyr

Conflicts with tidy packages ---------------------------------------------------
-

filter(): dplyr, stats
lag():    dplyr, stats
```

In [41]:
```
hp2001Q1 <- filter(housing, Date == 2001.25)
ggplot(hp2001Q1,
       aes(y = Structure.Cost, x = Land.Value)) +
geom_point()
```

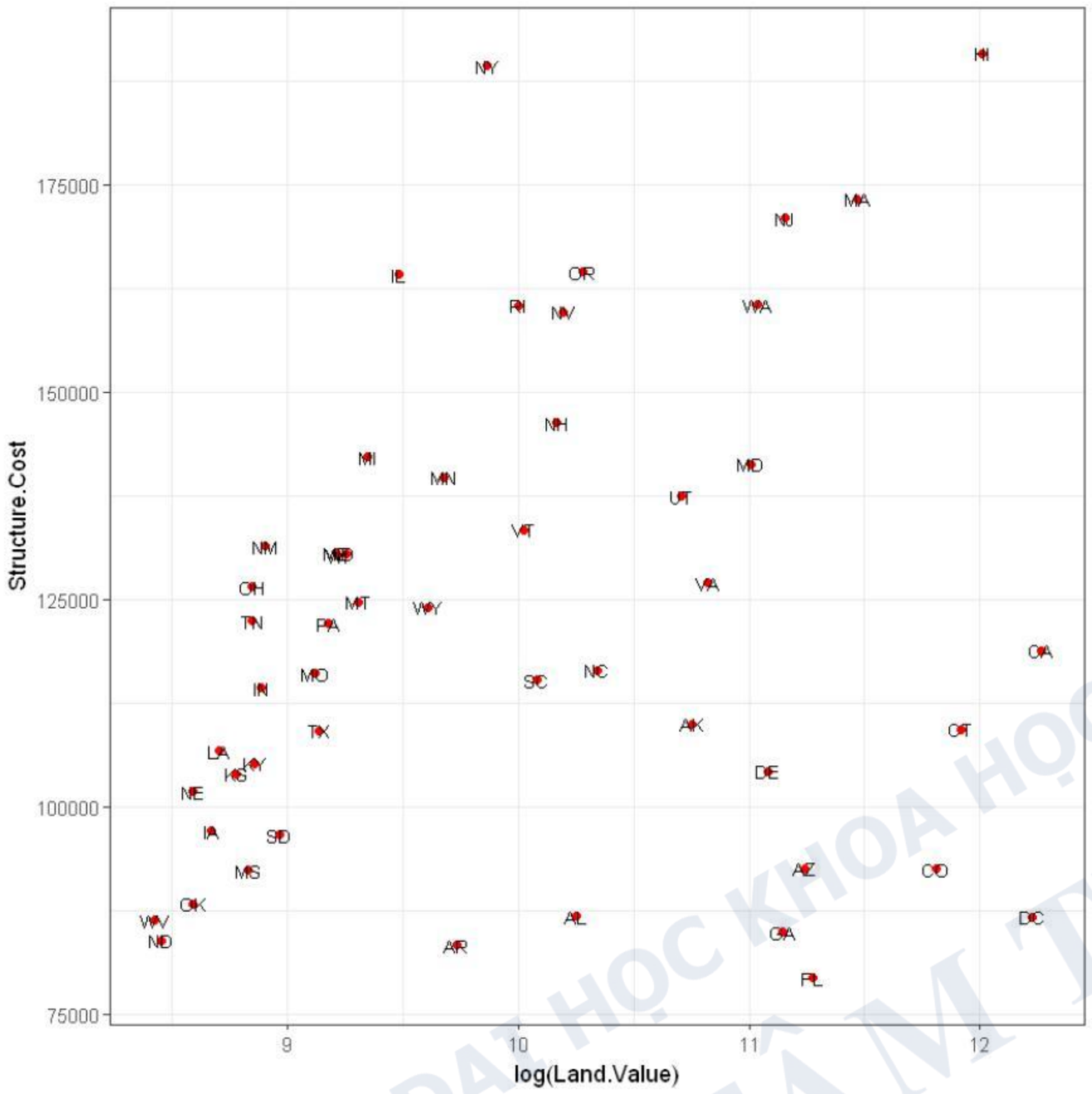


In [48]:
```
p1 = ggplot(hp2001Q1,
            aes(y = Structure.Cost, x = log(Land.Value))) +
  geom_point(color="red")
```

In [49]:
```
p1
```

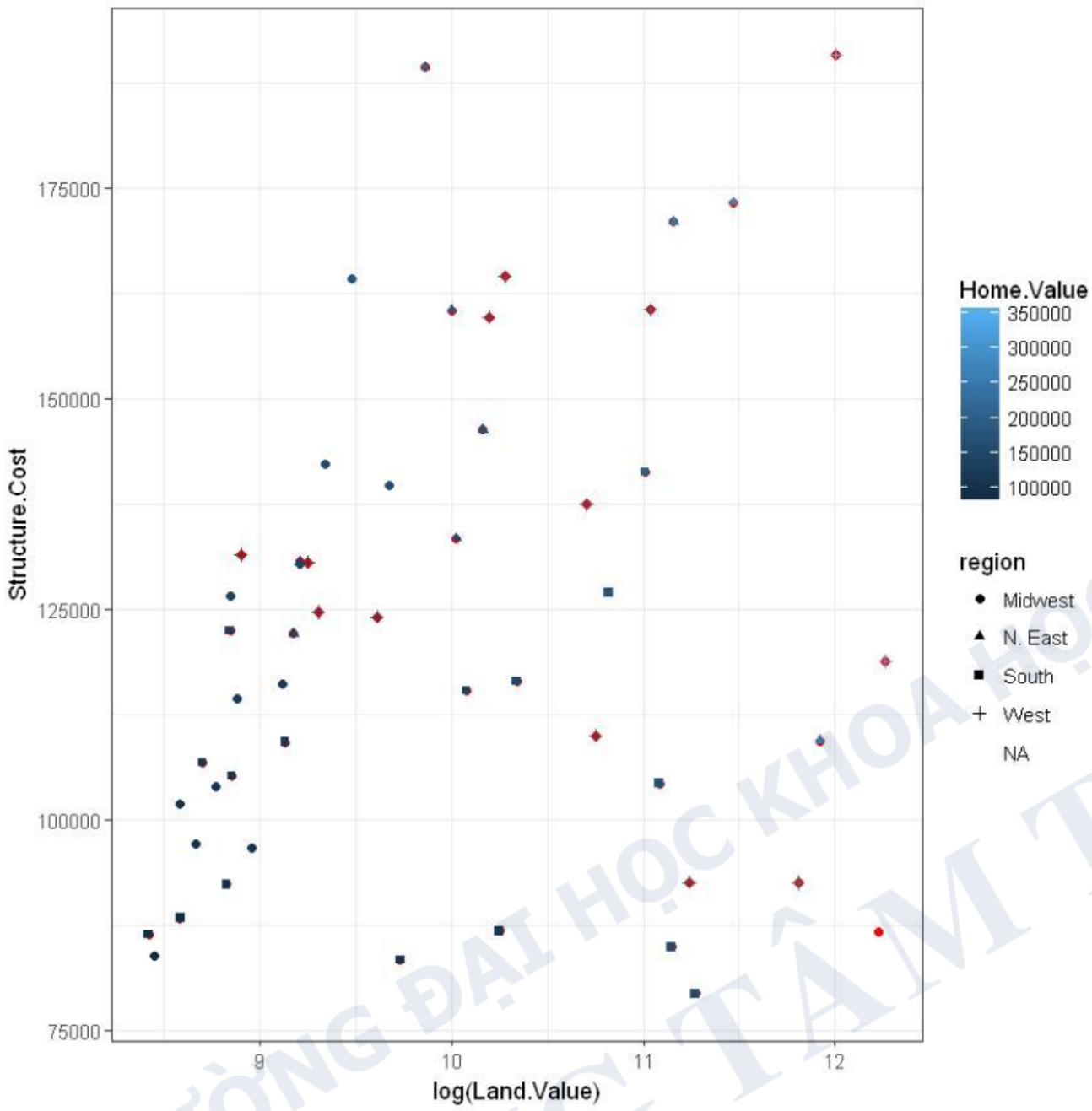


In [50]:
```
p1 +
  geom_text(aes(label=State), size = 3)
```



In [51]:
```
p1 +
  geom_point(aes(color=Home.Value, shape = region))
```

```
Warning message:
"Removed 1 rows containing missing values (geom_point)."
```



Exercise 3: EconomistData

In [52]:
```
dat <- read.csv("EconomistData.csv")
```

In [53]:
```
print(head(dat))
```

```
      Country HDI.Rank   HDI CPI            Region
1 Afghanistan      172 0.398 1.5      Asia Pacific
2     Albania       70 0.739 3.1 East EU Cemt Asia
3     Algeria       96 0.698 2.9              MENA
4      Angola      148 0.486 2.0               SSA
5   Argentina       45 0.797 3.0          Americas
6     Armenia       86 0.716 2.6 East EU Cemt Asia
```

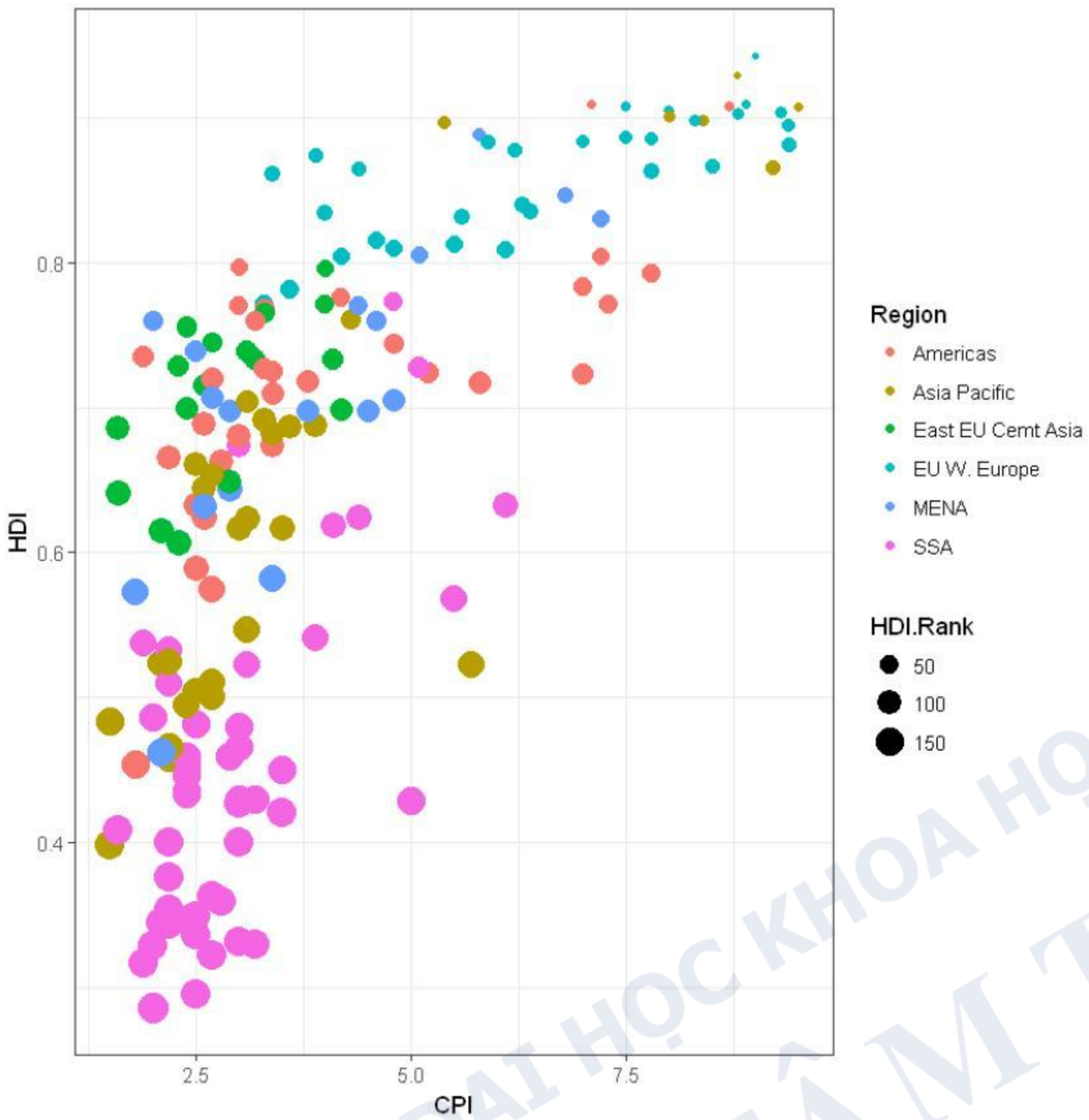
In [54]:
```
summary(dat)
```

```
       Country       HDI.Rank          HDI              CPI       
 Afghanistan:  1   Min.   :  1.00   Min.   :0.2860   Min.   :1.500  
 Albania    :  1   1st Qu.: 47.00   1st Qu.:0.5090   1st Qu.:2.500  
 Algeria    :  1   Median : 96.00   Median :0.6980   Median :3.200  
 Angola     :  1   Mean   : 95.28   Mean   :0.6581   Mean   :4.052  
 Argentina  :  1   3rd Qu.:143.00   3rd Qu.:0.7930   3rd Qu.:5.100  
 Armenia    :  1   Max.   :187.00   Max.   :0.9430   Max.   :9.500  
 (Other)    :167                                                    
               Region
 Americas         :31
 Asia Pacific     :30
 East EU Cemt Asia:18
 EU W. Europe     :30
 MENA             :18
 SSA              :46
```

In [55]:
```
str(dat)
```

```
'data.frame':   173 obs. of  5 variables:
 $ Country : Factor w/ 173 levels "Afghanistan",..: 1 2 3 4 5 6 7 8 9 10 ...
 $ HDI.Rank: int  172 70 96 148 45 86 2 19 91 53 ...
 $ HDI     : num  0.398 0.739 0.698 0.486 0.797 0.716 0.929 0.885 0.7 0.771 ...
 $ CPI     : num  1.5 3.1 2.9 2 3 2.6 8.8 7.8 2.4 7.3 ...
 $ Region  : Factor w/ 6 levels "Americas","Asia Pacific",..: 2 3 5 6 1 3 2 4 3
1 ...
```

In [56]:
```
ggplot(dat, aes(x = CPI, y = HDI)) +
    geom_point(aes(color = Region, size =  HDI.Rank))
```



In [57]:
```
ggplot(dat, aes(x = CPI, y = HDI)) +
  geom_point() +
  geom_smooth(method = "lm")
```