STA 445

CONSTANT YAOKUMAH

2023-11-15

```
knitr::opts_chunk$set(echo = TRUE, warning = FALSE, message = FALSE)
```

```
library(ggplot2)
library(dplyr)
library(datasets)
library(tidyr)
library(ggrepel)
library(latex2exp)
```

1a.

The `rownames()` of the table gives the country names and you should create a new column that contains the country names. *`rownames`

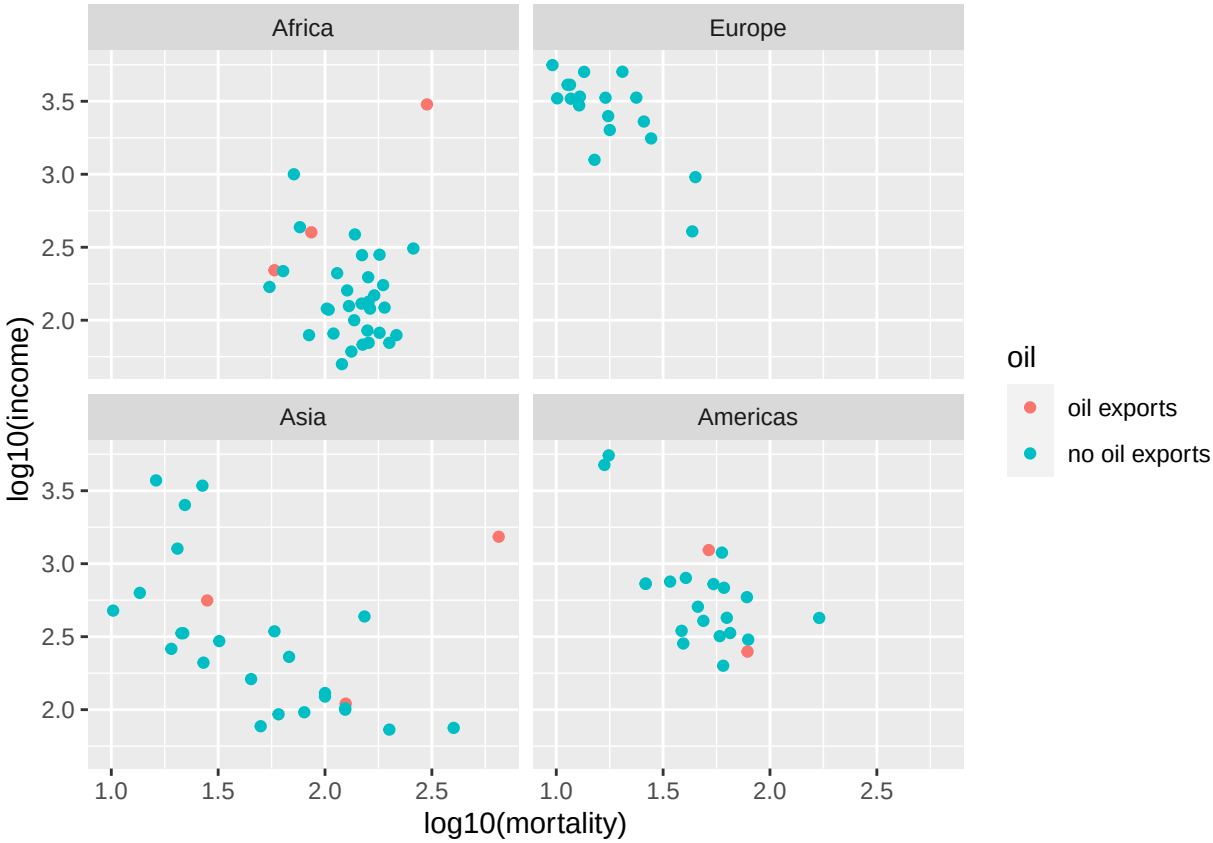
```
data('infmort', package = 'faraway')

infmort_data <- infmort %>%
  mutate(Country = row.names(infmort))
infmort_data <- drop_na(infmort_data)
head(infmort_data)
```

```
##                       region income mortality            oil
## Australia               Asia   3426      26.7 no oil exports
## Austria               Europe   3350      23.7 no oil exports
## Belgium               Europe   3346      17.0 no oil exports
## Canada              Americas   4751      16.8 no oil exports
## Denmark               Europe   5029      13.5 no oil exports
## Finland               Europe   3312      10.1 no oil exports
##                              Country
## Australia           Australia
## Austria             Austria
## Belgium             Belgium
## Canada              Canada
## Denmark             Denmark
## Finland             Finland
```

1b

Create scatter plots with the `log10()` transformation inside the `aes()`command.

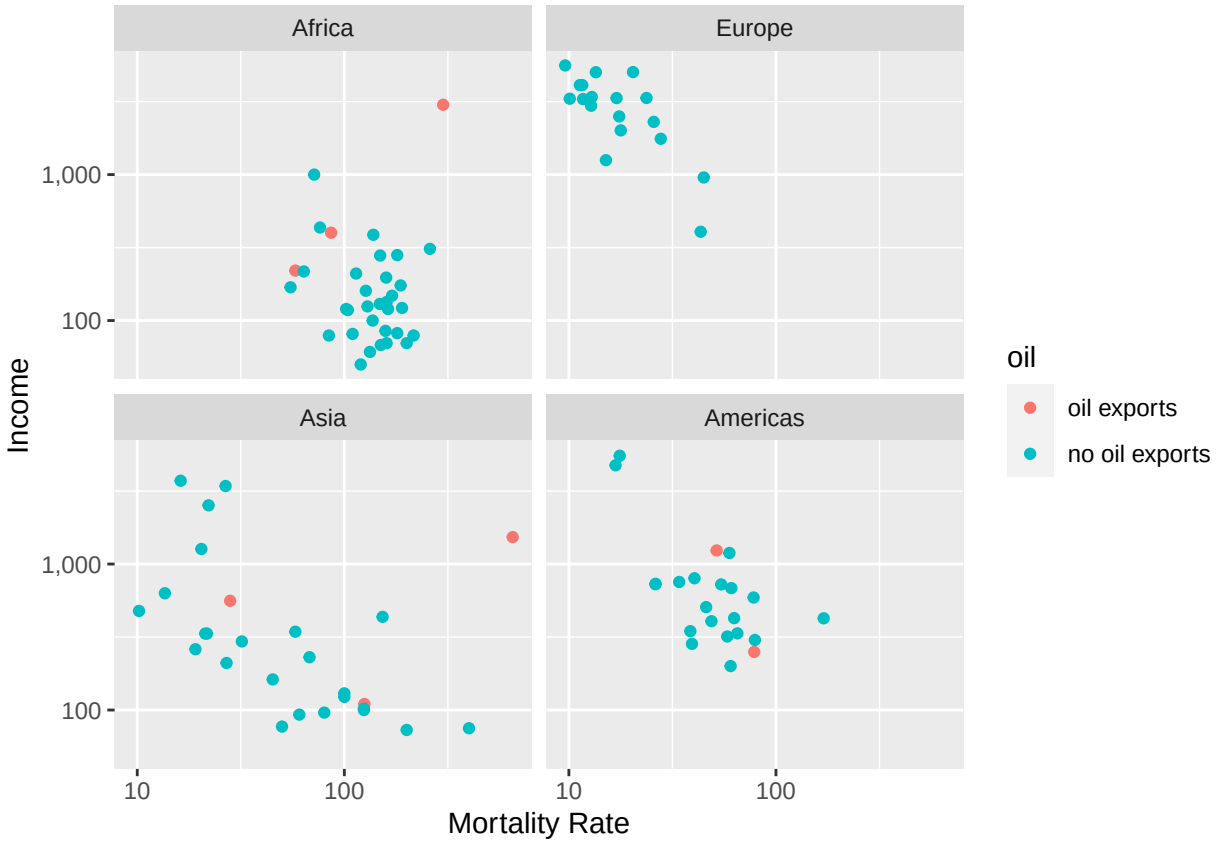
```
ggplot(infmort_data, aes(x = log10(mortality), y = log10(income), color = oil)) +
  geom_point() + facet_wrap(.~region)
```



1c

Create the scatter plots using the `scale_x_log10()` and `scale_y_log10()`. Set the major and minor breaks to be useful and aesthetically pleasing. Comment on which version you find easier to read.

```
p2 <- ggplot(infmort_data, aes(x =mortality , y = income)) +
  geom_point(aes(color = oil)) +
  facet_wrap(.~ region) +
  labs(x = "Mortality Rate", y = " Income") +
  scale_x_log10(breaks = c(10, 100, 1000, 10000),labels = c("10", "100", "1,000", "10,000")) +
  scale_y_log10(breaks = c(100, 1000, 10000),labels = c("100", "1,000", "10,000"))

p2
```

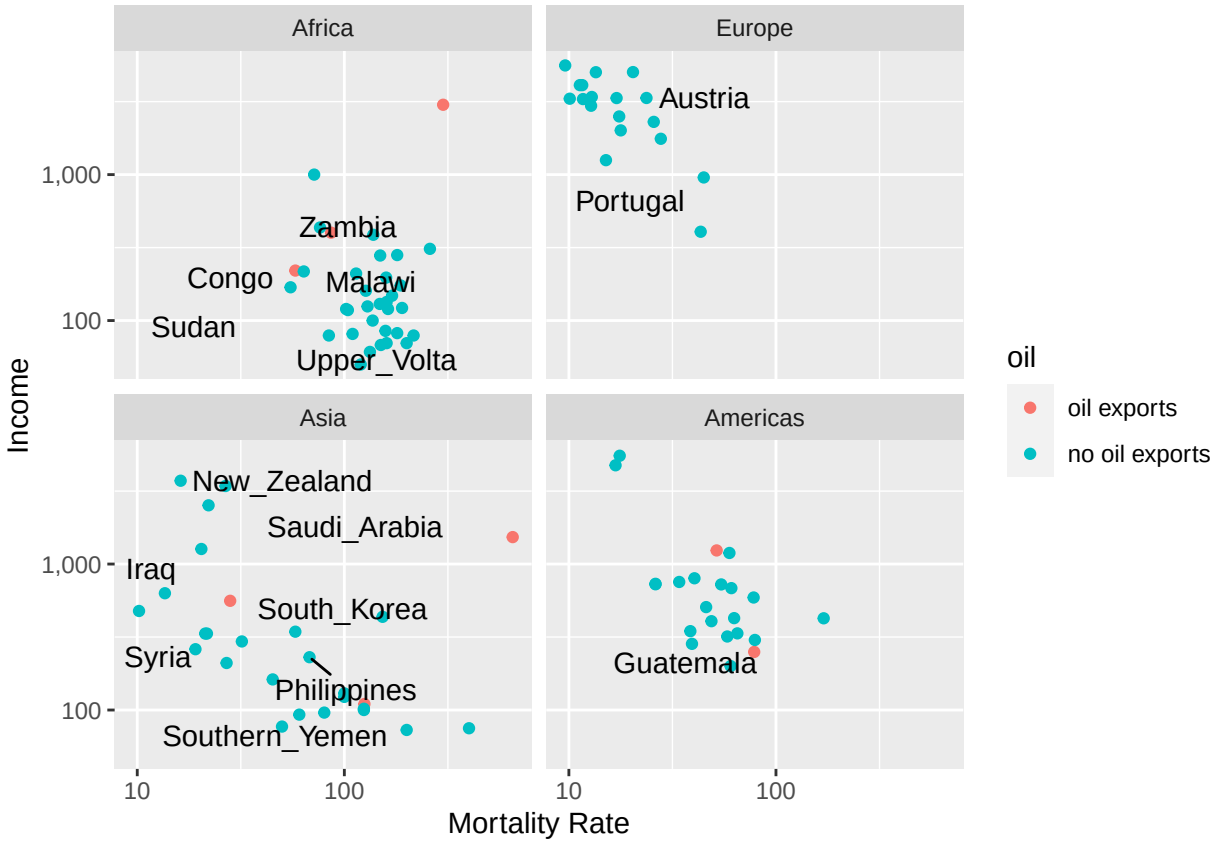


In general, using the scale_x_log10() and scale_y_log10() functions is preferred because it provides better control over the axis scales and breaks. The labels on the axes will be in the original units, and it's easier to interpret the data.

1d

Select 10-15 countries to label and do so using the geom_text_repel() function.

```
selected_countries <- sample(infmort_data$Country, 15)  # Select 10-15 countries to label

p3 <- p2 +
  geom_text_repel(data = subset(infmort_data, Country %in% selected_countries),
                  aes(label = Country))

p3
```



2a

Create a regression model for $y =$ `Volume` as a function of $x =$ `Height`.

```
data(trees)
model <- lm(Volume ~ Height, data = trees)
data <- trees %>% mutate(fit = fitted(model))
head(data)
```

```
##   Girth Height Volume      fit
## 1   8.3     70   10.3 20.91087
## 2   8.6     65   10.3 13.19412
## 3   8.8     63   10.2 10.10742
## 4  10.5     72   16.4 23.99757
## 5  10.7     81   18.8 37.88772
## 6  10.8     83   19.7 40.97442
```

2b

Using the `summary` command, get the y-intercept and slope of the regression line.

```
summary(model)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

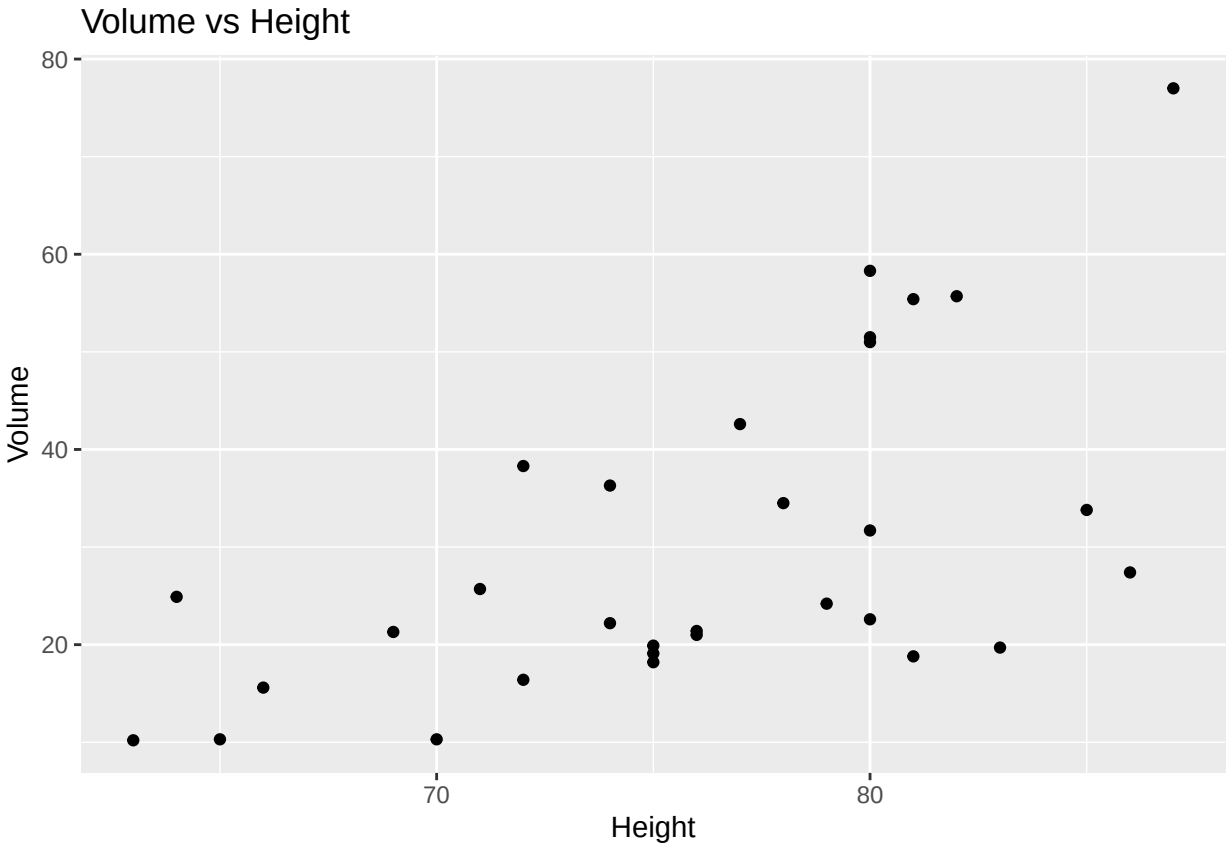
```
coef(model)
```

```
## (Intercept)      Height
##   -87.12361     1.54335
```

2c

Using ggplot2, create a scatter plot of Volume vs Height.

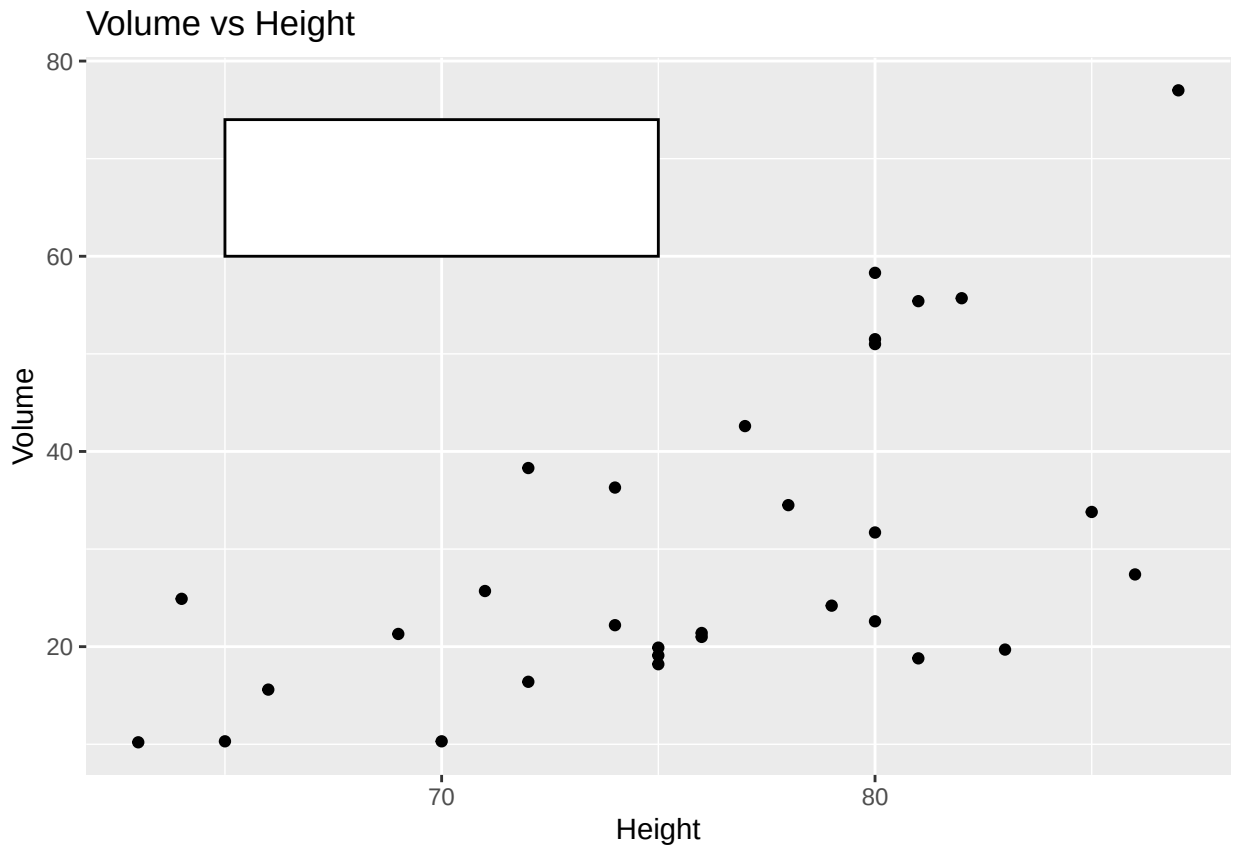
```
library(ggplot2)

p <- ggplot(data, aes(x = Height, y = Volume)) +
  geom_point() +
  labs(x = "Height", y = "Volume", title = "Volume vs Height")
p
```



2d

Create a nice white filled rectangle to add text information to using by adding the following annotation layer

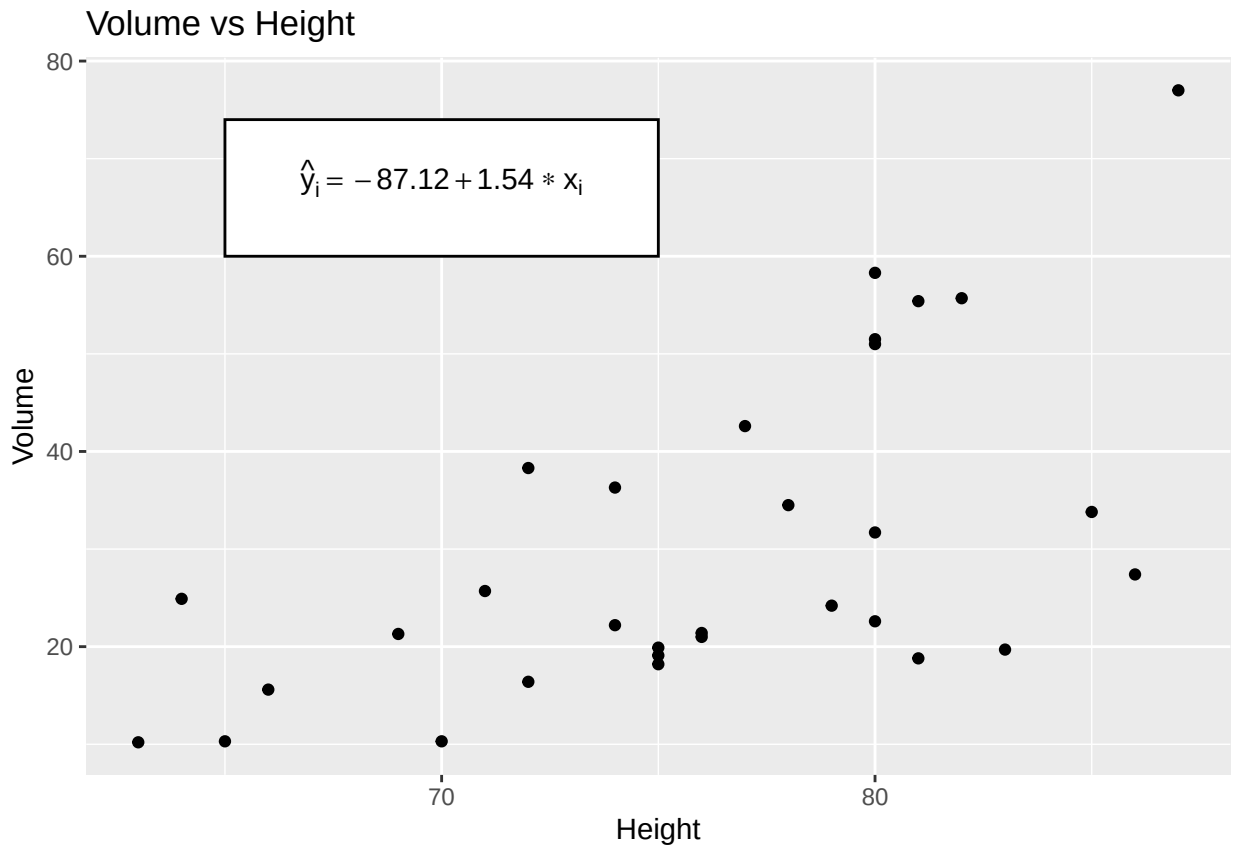
```
p1 <- p + annotate(
  "rect", xmin = 65, xmax = 75, ymin = 60, ymax = 74,
  fill = "white", color = "black"
)
p1
```



2e

Add some annotation text to write the equation of the line

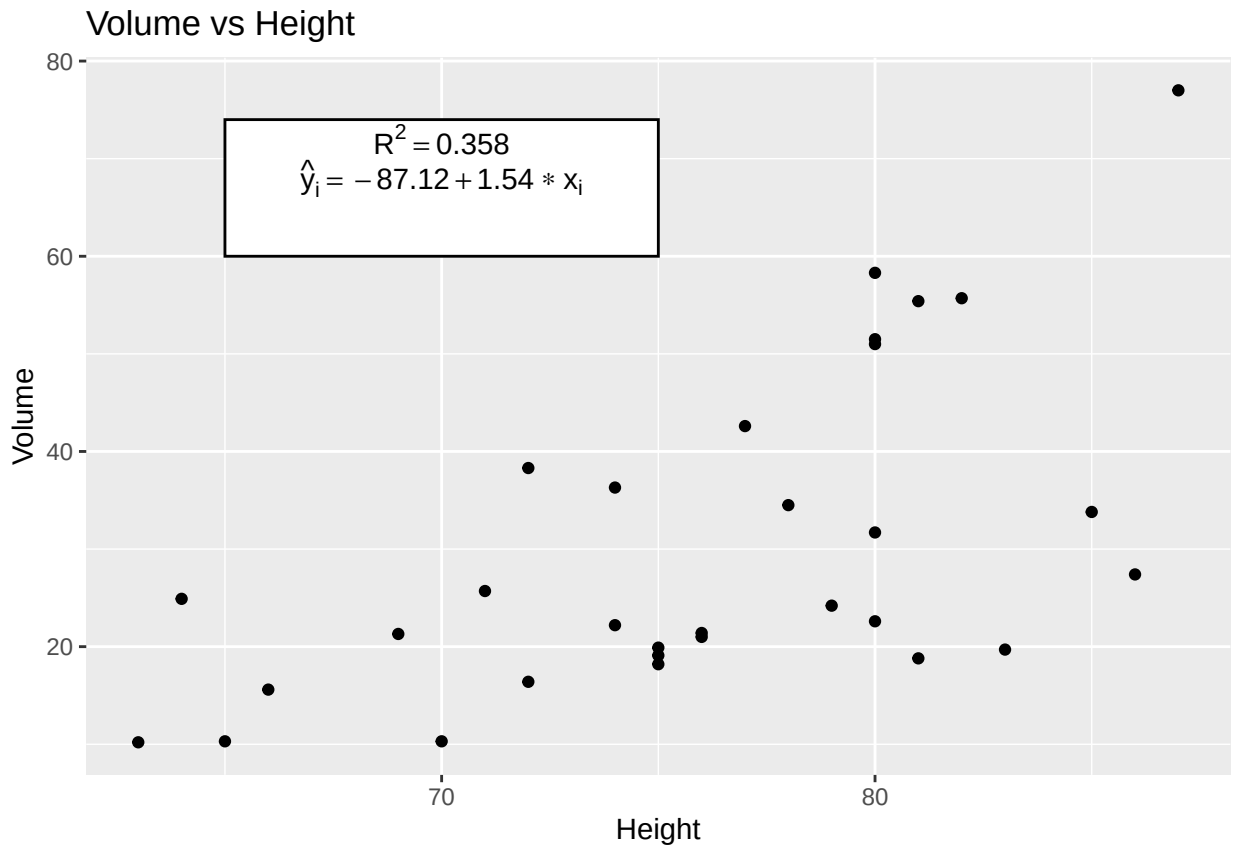
```
p2 <- p1 + annotate(
  "text", x = 70, y = 68,
  label = latex2exp::TeX('$\\hat{y}_{i} = -87.12 + 1.54 * x_{i}$'),

)
p2
```



2f

Add annotation to add $R^2 = 0.358$

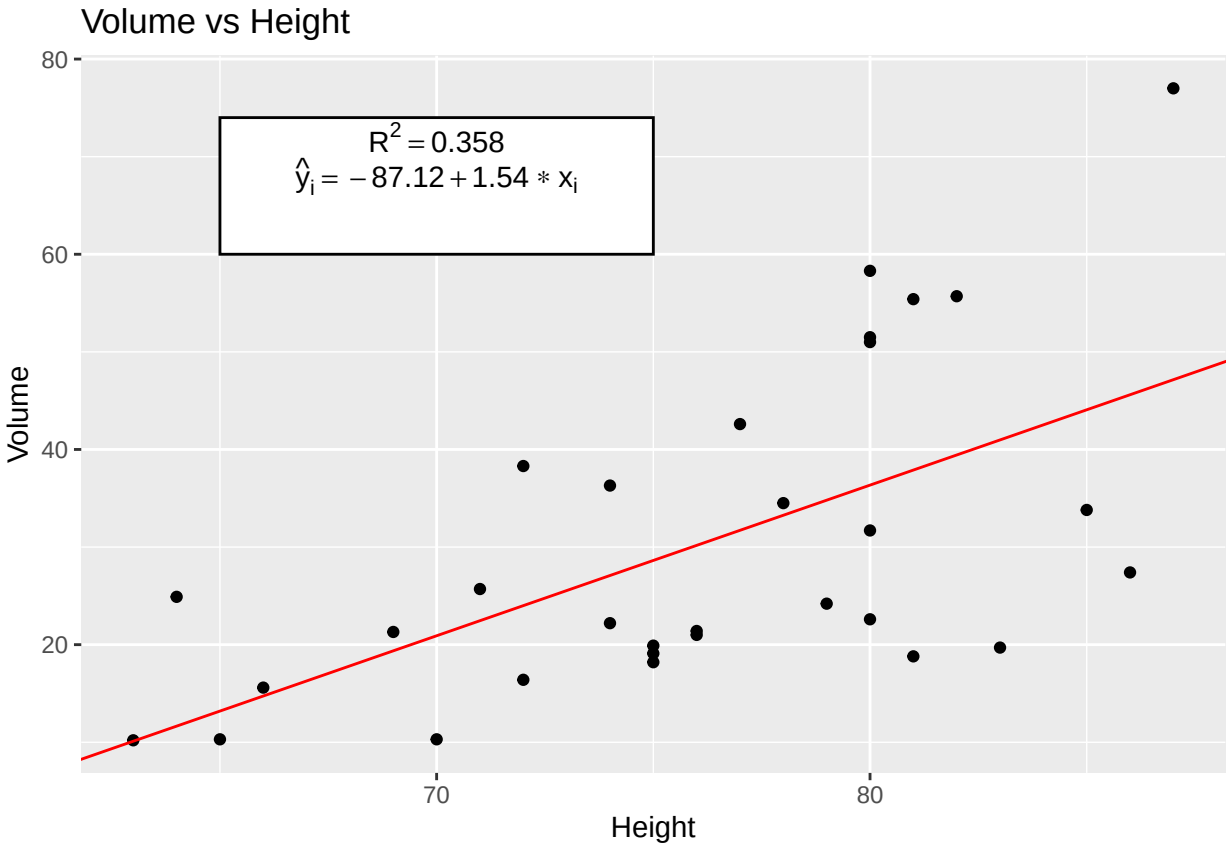
```
p3 <- p2 + annotate(
  "text", x = 70, y = 72,
  label = latex2exp('$R^{2} = 0.358$')
)
p3
```



2g

Add the regression line in red. The most convenient layer function to uses is `geom_abline()`

```
p4 <- p3 + geom_abline(intercept = coef(model)[1], slope = coef(model)[2], color = "red")
p4
```


Volume vs Height
$R^2 = 0.358$
$\hat{y}_i = -87.12 + 1.54 * x_i$
Volume
Height
80
60
40
20
70
80