

# Assignment5

Kyle Bambling

2023-11-14

## Question 1

The `infmort` data set from the package `faraway` gives the infant mortality rate for a variety of countries. The information is relatively out of date (from 1970s?), but will be fun to graph. Visualize the data using by creating scatter plots of mortality vs income while faceting using `region` and setting color by `oil` export status. Utilize a  $\log_{10}$  transformation for both `mortality` and `income` axes. This can be done either by doing the transformation inside the `aes()` command or by utilizing the `scale_x_log10()` or `scale_y_log10()` layers. The critical difference is if the scales are on the original vs log transformed scale. Experiment with both and see which you prefer.

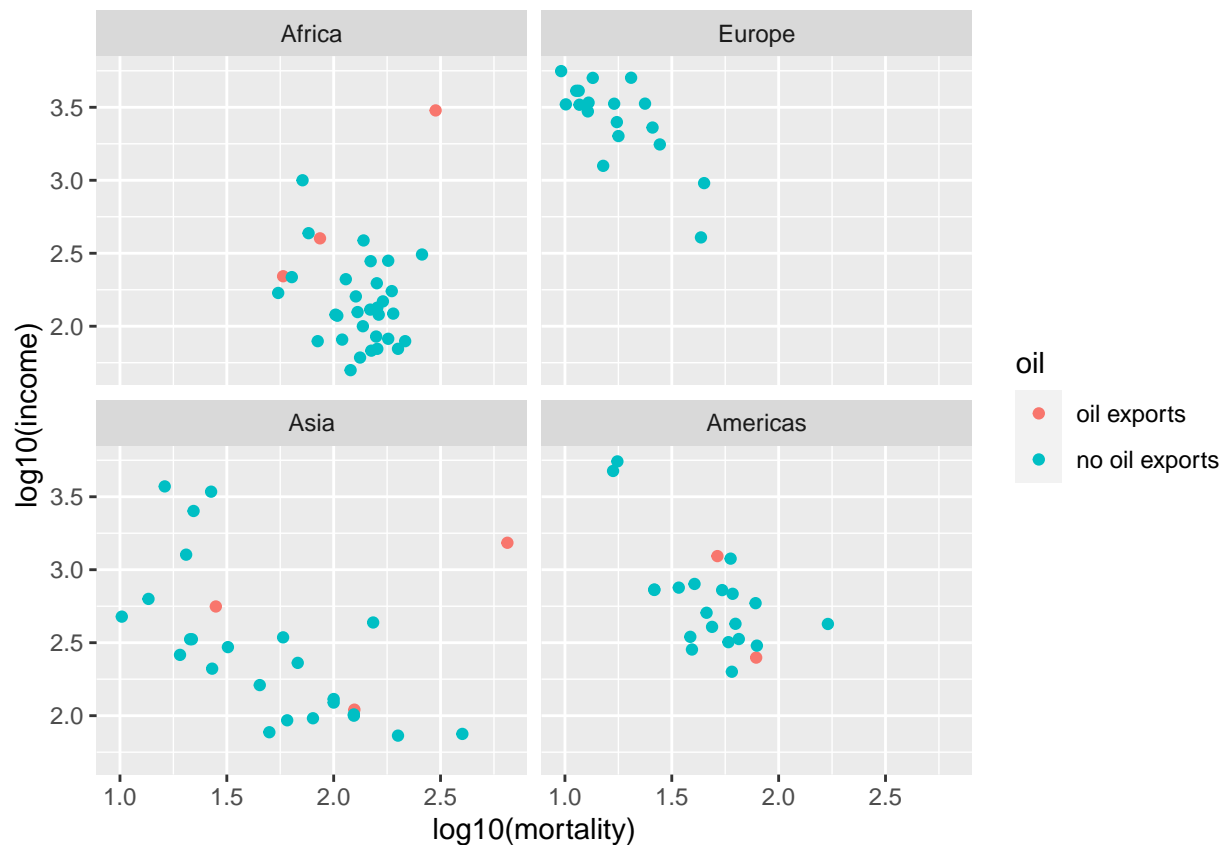
```
data(infmort, package = 'faraway')
infmort <- infmort %>% drop_na()
```

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

```
infmort <- infmort %>%
  mutate(country = rownames(infmort))
```

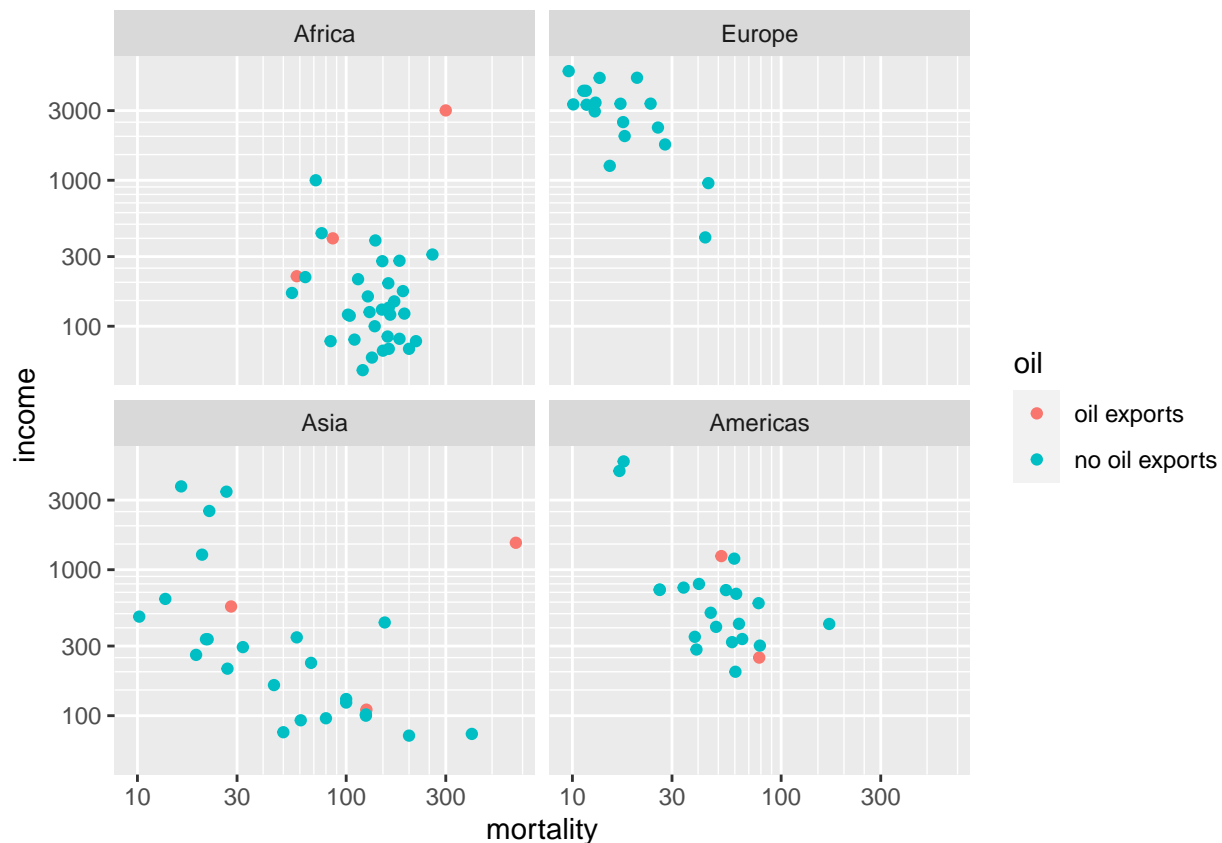
- b) Create scatter plots with the `log10()` transformation inside the `aes()` command.

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



c) 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.

```
ggplot(infmort, aes(x=mortality, y=income, color=oil)) +
  geom_point() +
  facet_wrap(~ region) +
  scale_x_log10(minor=c(seq(30,100, 10), seq(100,300, 50), seq(300,1000, 200))) +
  scale_y_log10(minor=c(seq(100,300, 50), seq(300,1000, 100), seq(1000,3000, 500)))
```



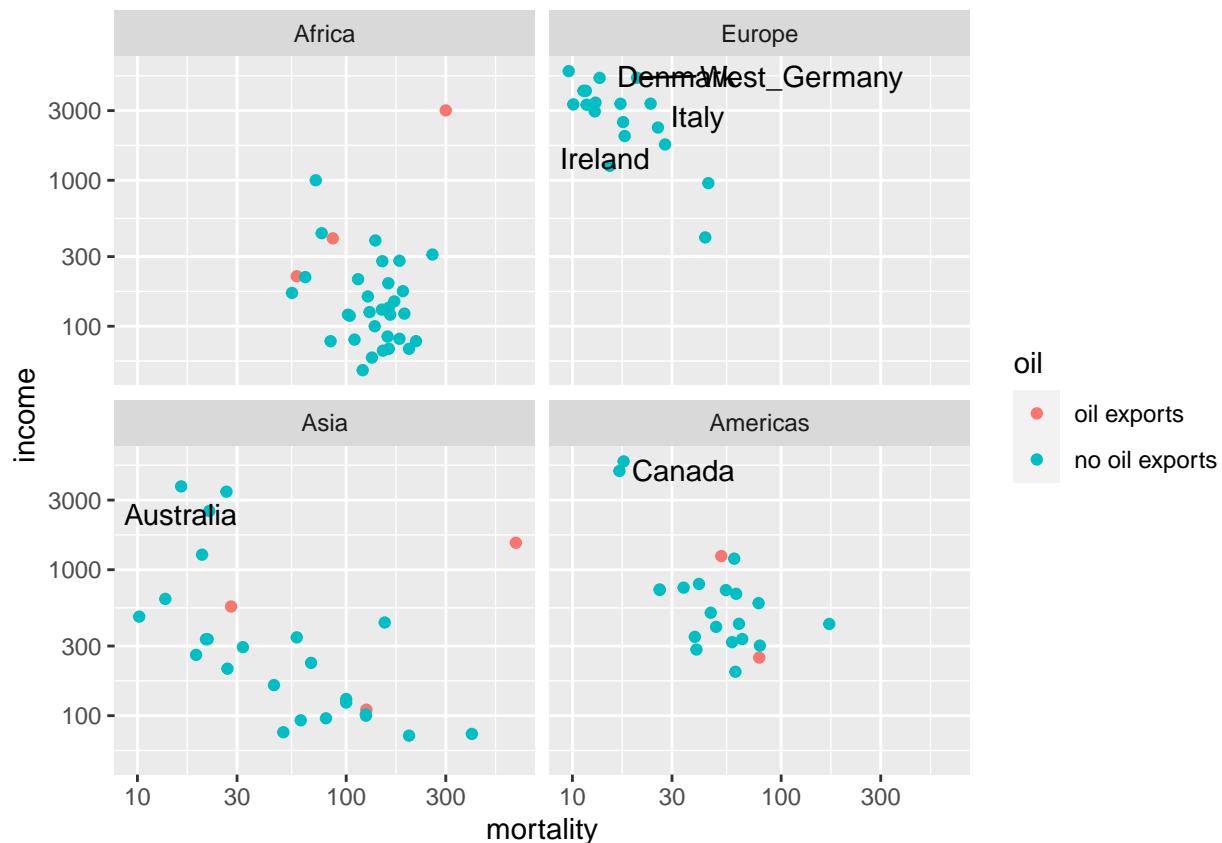
d) The package `ggrepel` contains functions `geom_text_repel()` and `geom_label_repel()` that mimic the basic `geom_text()` and `geom_label()` functions in `ggplot2`, but work to make sure the labels don't overlap. Select 10-15 countries to label and do so using the `geom_text_repel()` function.

```
infmort <- infmort %>% mutate(sl.country = ifelse(country==country[1:10], country, NA))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `sl.country = ifelse(country == country[1:10], country, NA)`:
## Caused by warning in `country == country[1:10]`:
## ! longer object length is not a multiple of shorter object length
```

```
ggplot(infmort, aes(x=mortality, y=income)) +
  geom_point(aes(color=oil)) +
  facet_wrap(~ region) +
  scale_x_log10() +
  scale_y_log10() +
  ggrepel::geom_text_repel(aes(label = sl.country))
```

```
## Warning: Removed 91 rows containing missing values (`geom_text_repel()`).
## Warning: ggrepel: 4 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



## Question 2

Using the `datasets::trees` data, complete the following:

- a) Create a regression model for  $y = \text{Volume}$  as a function of  $x = \text{Height}$ .

```
data(trees, package = 'datasets')
trees.reg <- lm(Volume ~ Height, data = trees)
trees <- trees %>% mutate(fit = fitted(trees.reg))
```

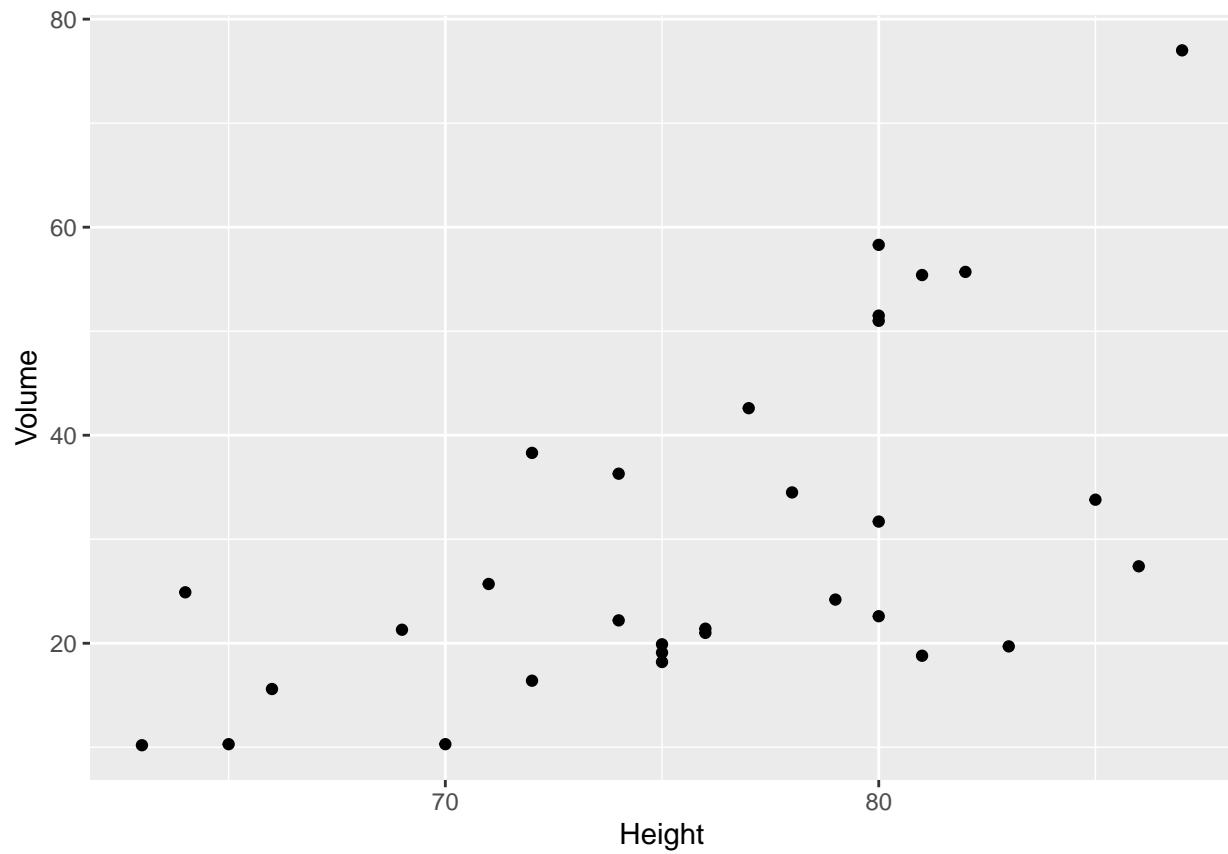
- b) Using the `summary` command, get the y-intercept and slope of the regression line.

```
trees.summ <- tidy(trees.reg)
trees.summ

## # A tibble: 2 x 5
##   term      estimate std.error statistic  p.value
##   <chr>      <dbl>    <dbl>     <dbl>   <dbl>
## 1 (Intercept) -87.1      29.3     -2.98 0.00583
## 2 Height       1.54     0.384      4.02 0.000378
```

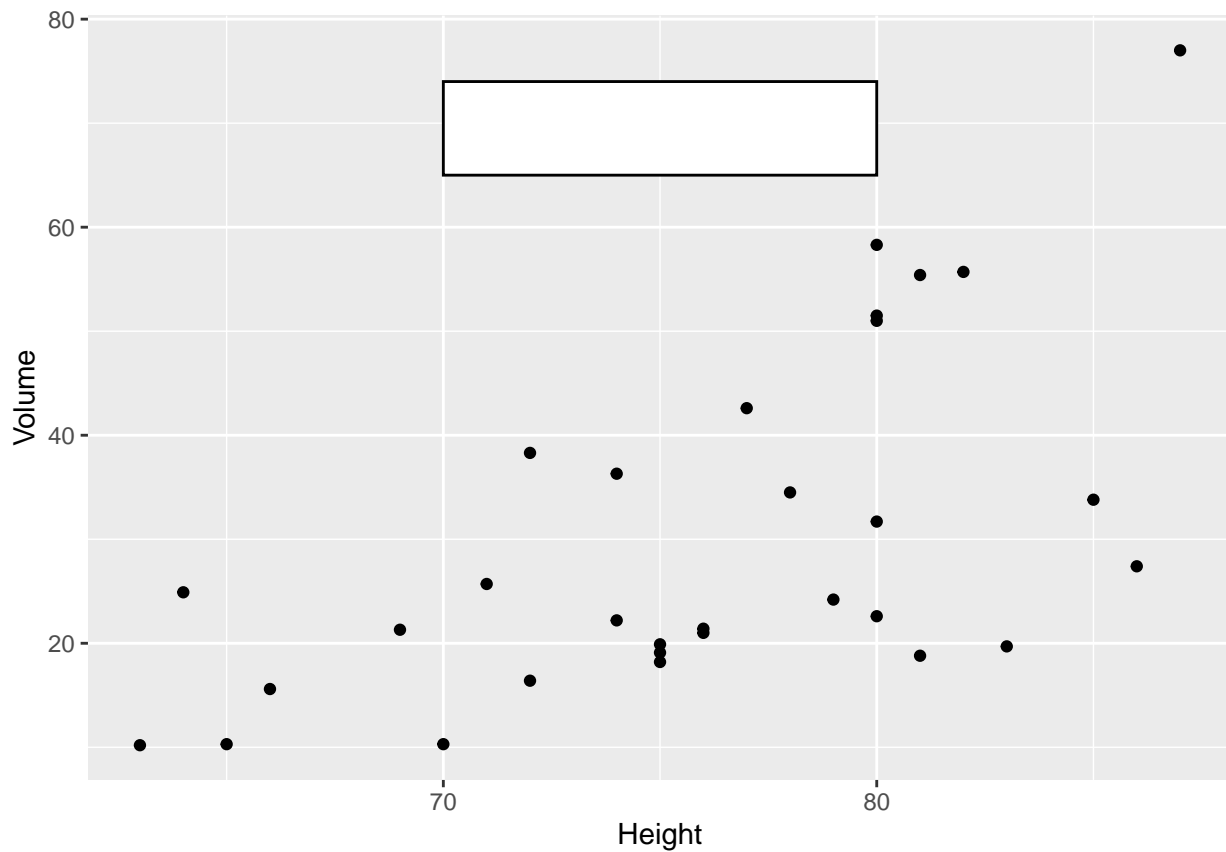
- c) Using `ggplot2`, create a scatter plot of Volume vs Height.

```
ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point()
```



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

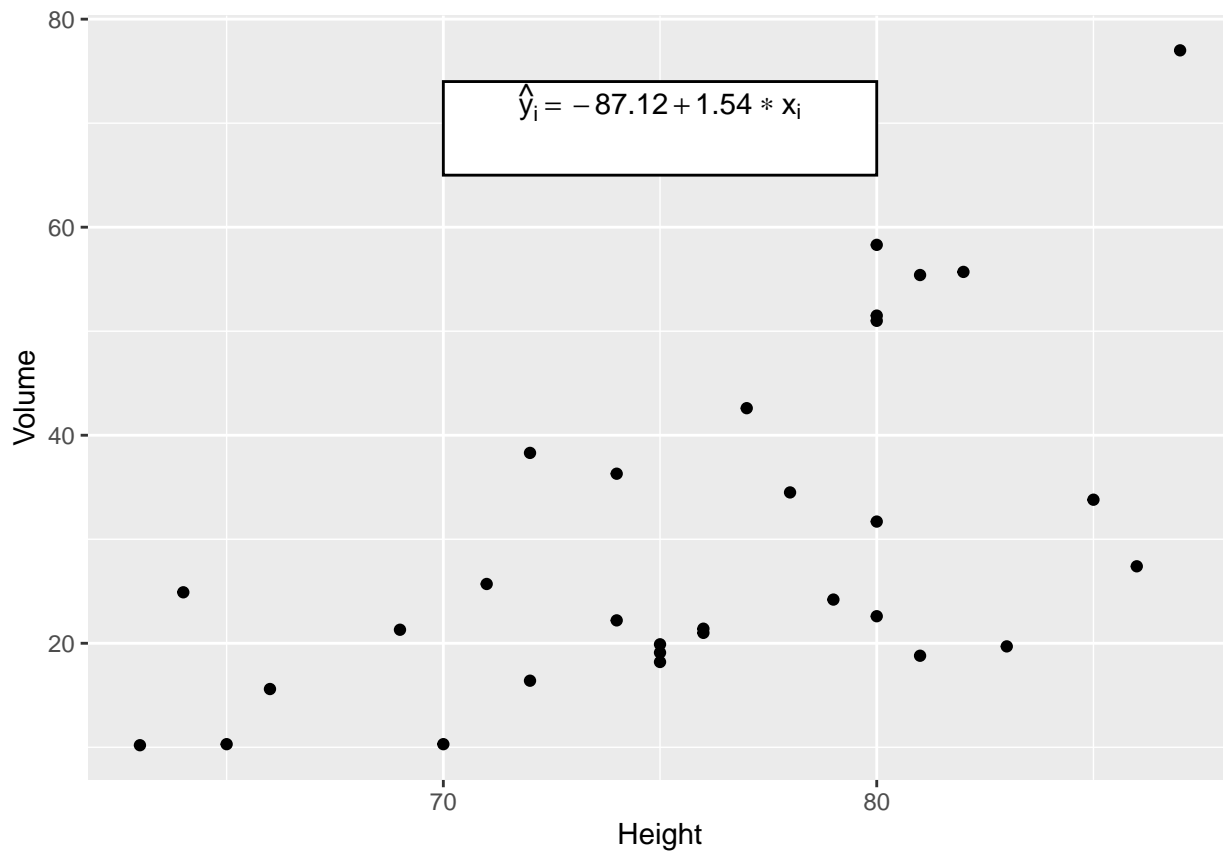
```
ggplot(trees, aes(x=Height, y=Volume)) +  
  geom_point() +  
  annotate('rect', xmin=70, xmax=80, ymin=65, ymax=74,  
         fill='white', color='black')
```



e) Add some annotation text to write the equation of the line  $\hat{y}_i = -87.12 + 1.54 * x_i$  in the text area.

```
ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point() +
  annotate('rect', xmin=70, xmax=80, ymin=65, ymax=74,
         fill='white', color='black') +
  annotate('text', x = 75, y = 72, label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$'))

## Warning in is.na(x): is.na() applied to non-(list or vector) of type
## 'expression'
```

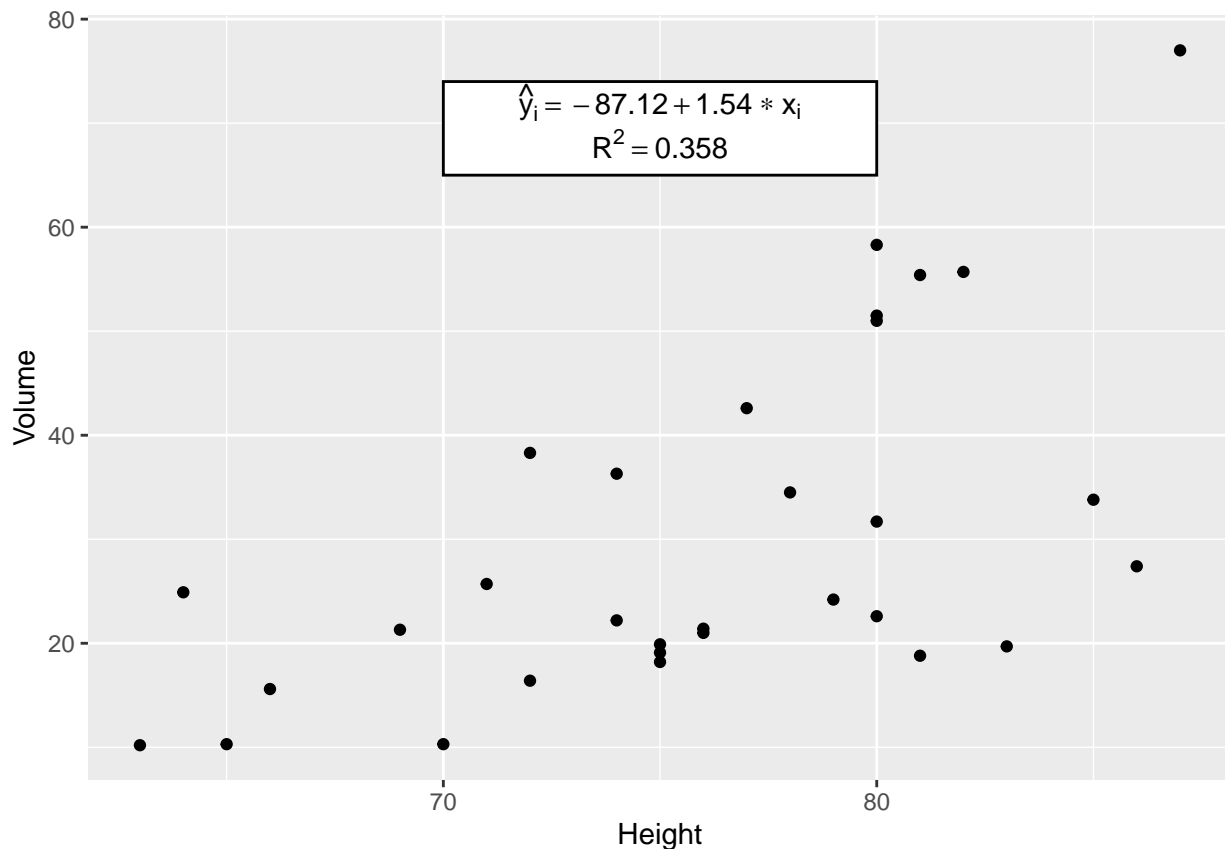


f) Add annotation to add  $R^2 = 0.358$

```
ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point() +
  annotate('rect', xmin=70, xmax=80, ymin=65, ymax=74,
         fill='white', color='black') +
  annotate('text', x = 75, y = 72, label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$')) +
  annotate('text', x = 75, y = 68, label=latex2exp::TeX('$R^2 = 0.358$'))
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type
## 'expression'
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type
## 'expression'
```



g) Add the regression line in red. The most convenient layer function to use is `geom_abline()`. It appears that the `annotate` doesn't work with `geom_abline()` so you'll have to call it directly.

```
ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point() +
  annotate('rect', xmin=70, xmax=80, ymin=65, ymax=74,
         fill='white', color='black') +
  annotate('text', x = 75, y = 72, label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$')) +
  annotate('text', x = 75, y = 68, label=latex2exp::TeX('$R^2 = 0.358$')) +
  geom_line(aes(y = trees$fit), color = 'red')
```

```
## Warning: Use of `trees$fit` is discouraged.
```

```
## i Use `fit` instead.
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type
```

```
## 'expression'
```

```
## Warning in is.na(x): is.na() applied to non-(list or vector) of type
```

```
## 'expression'
```



