STA445 Assignment 5

Namiko Service

2023-11-09

Question 1

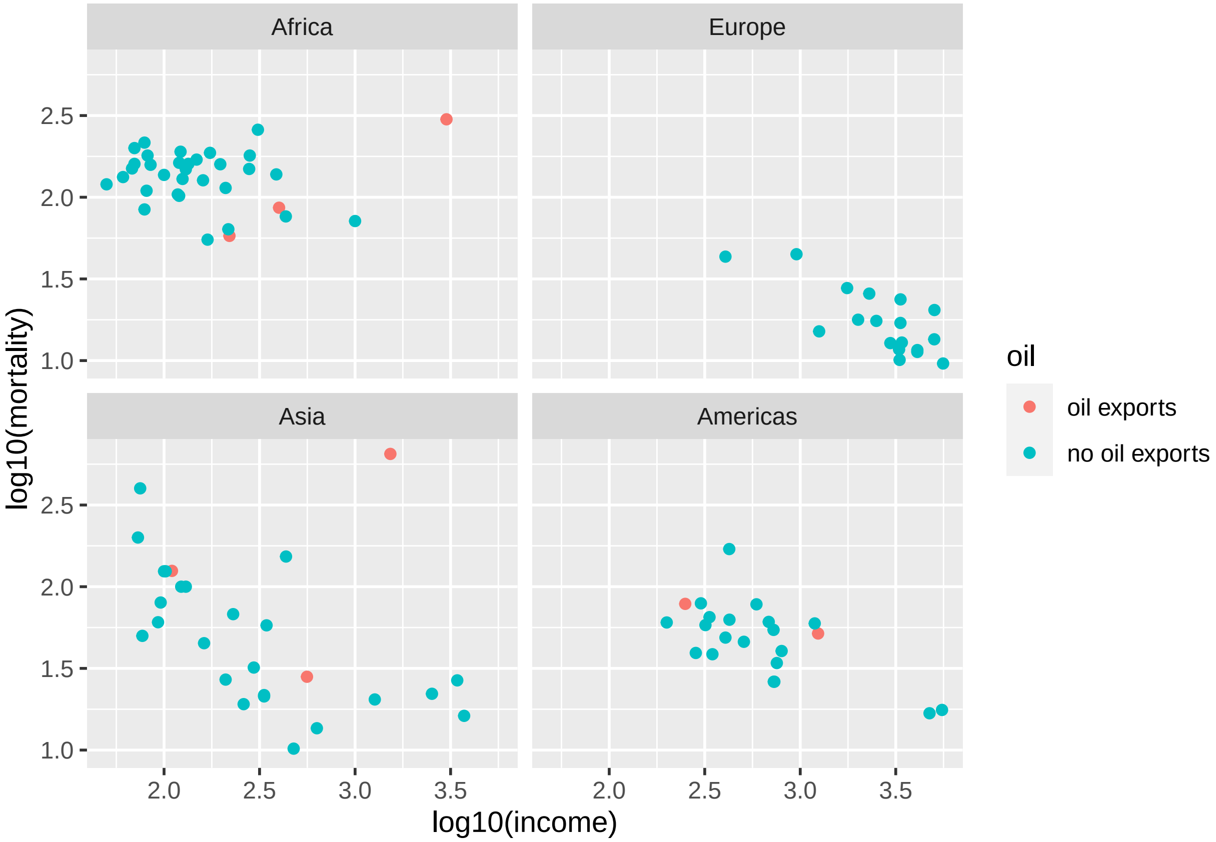
The `infmort` data set from the package `faraway` gives the infant mortality rate for a variety of countries. The information is relatively out of date (from 1970s?), but will be fun to graph. Visualize the data using by creating scatter plots of mortality vs income while faceting using `region` and 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.

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

```r
library(faraway)
data(infmort)

infmort1<-infmort%>%
  mutate(Country=rownames(infmort))%>%
  drop_na()%>%
  mutate(lable=ifelse(str_detect(string = Country, '^[ABab]' ), Country, ""))
```

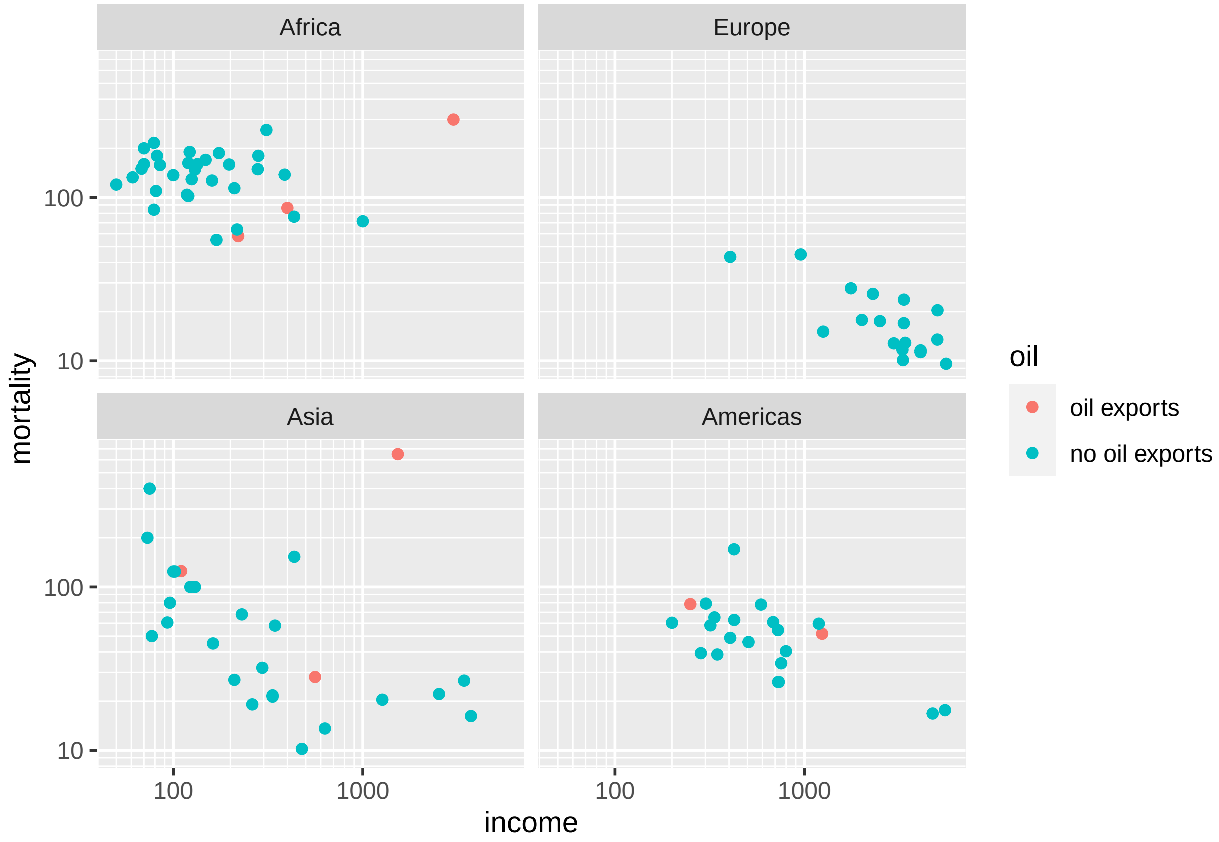
b) Create scatter plots with the `log10()` transformation inside the `aes()` command.

```r
ggplot(data = infmort1, aes(x=log10(income), y=log10(mortality)))+
  geom_point(aes(color=oil))+
  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.

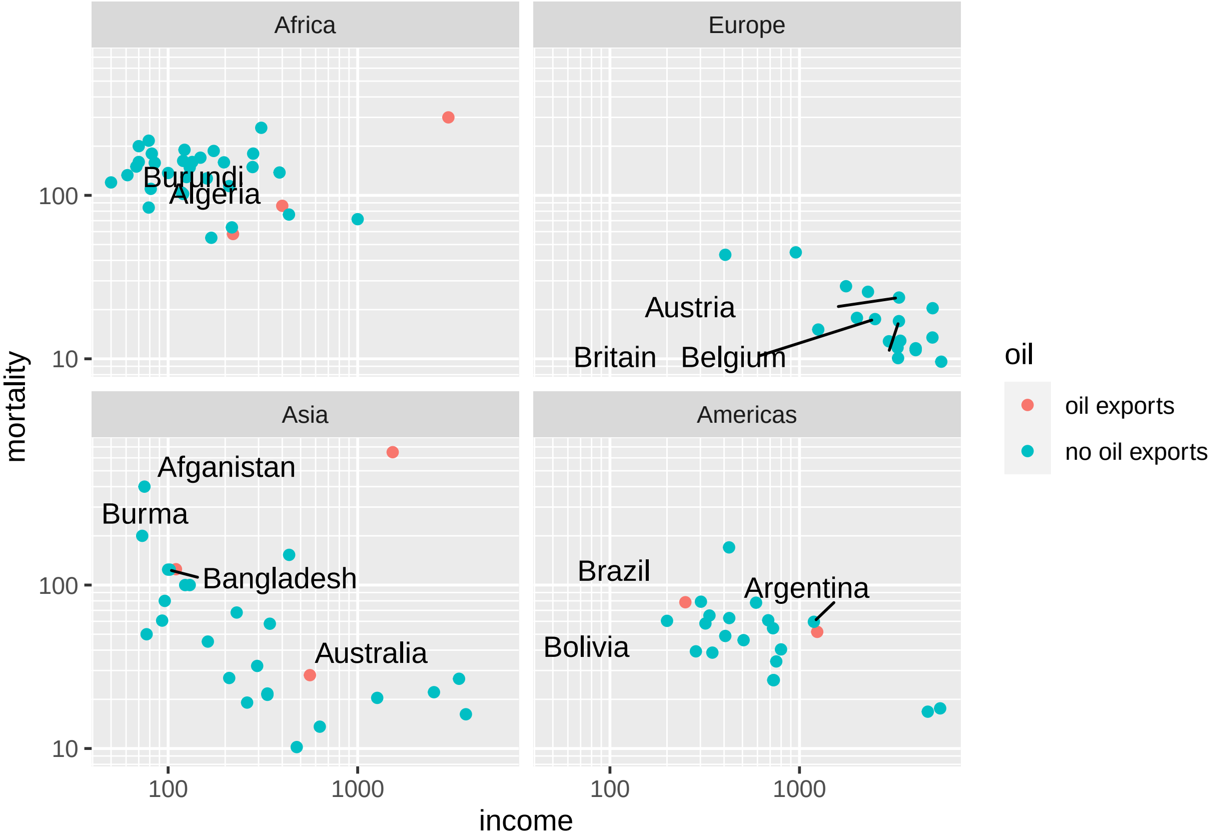
```
ggplot(data = infmort1, aes(x=income, y=mortality))+
  geom_point(aes(color=oil))+
  facet_wrap(~region)+
  scale_x_log10(breaks=c(1,10,100, 1000),
                minor=c(1:10,
                        seq( 10, 100,by=10 ),
                        seq(100,1000,by=100)))+
  scale_y_log10(breaks=c(1,10,100),
                minor=c(1:10,
                        seq( 10, 100,by=10 ),
                        seq(100,1000,by=100))
      )
```



I find the graph that uses the log_scale_x and log_scale_y to be easier to read because the untransformed axes make it easier to interpret the values.

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.

```
library(ggrepel)
ggplot(data = infmort1, aes(x=income, y=mortality, label=lable))+
  geom_point(aes(color=oil))+
  facet_wrap(~region)+
  scale_x_log10(breaks=c(1,10,100, 1000),
                minor=c(1:10,
                        seq( 10, 100,by=10 ),
                        seq(100,1000,by=100)))+
  scale_y_log10(breaks=c(1,10,100),
                minor=c(1:10,
                        seq( 10, 100,by=10 ),
                        seq(100,1000,by=100)))+
  geom_text_repel(max.overlaps = Inf)
```



Question 2

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

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

```
data(trees)
model <- lm( Volume ~ Height, data = trees )
```

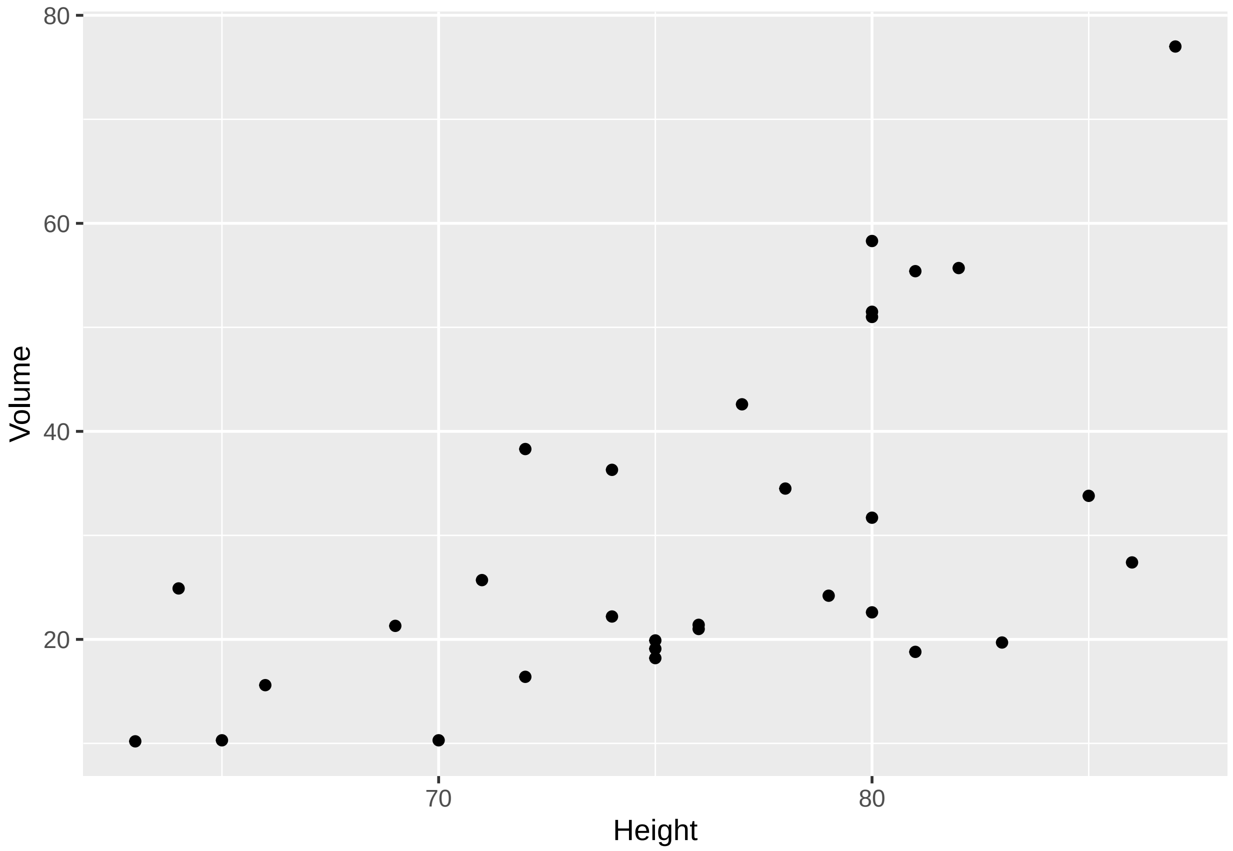
b) Using the `summary` command, get the y-intercept and slope of the regression line.

```
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
```

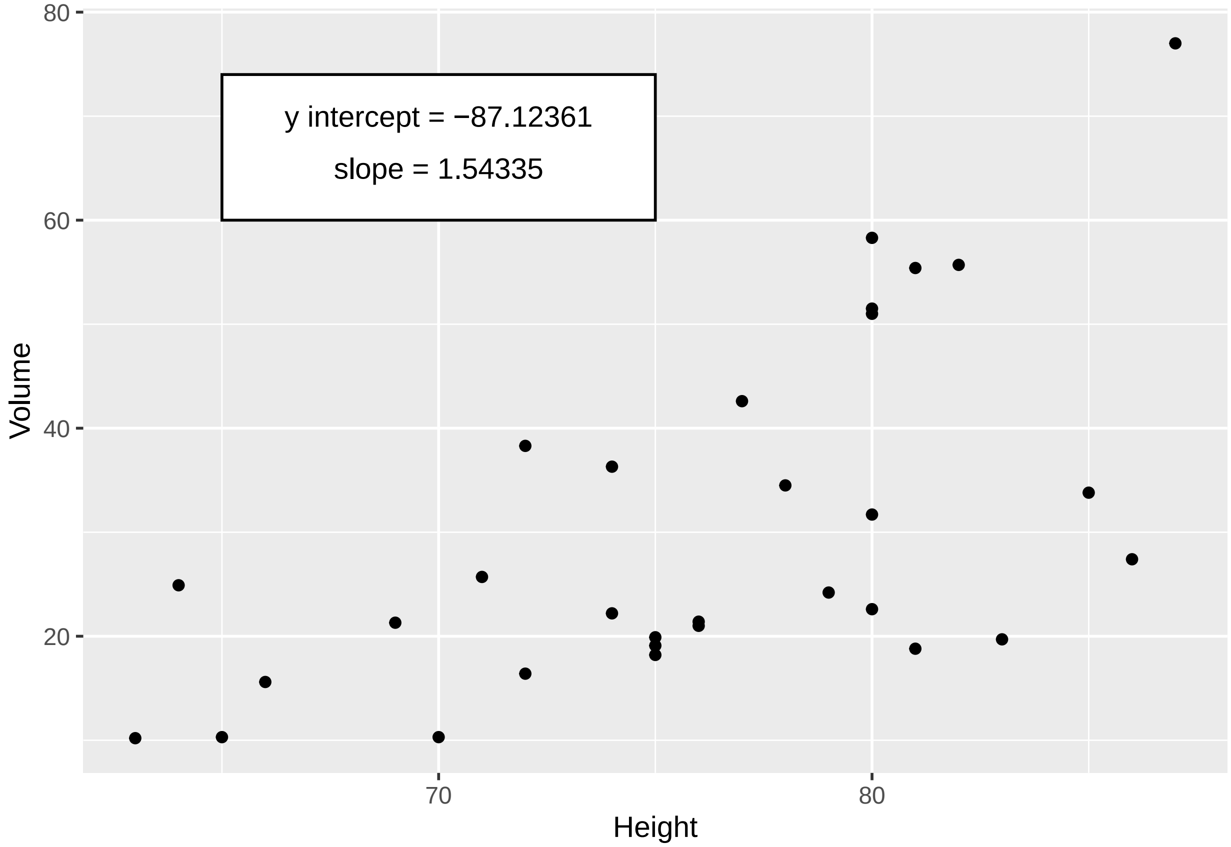
c) Using ggplot2, create a scatter plot of Volume vs Height.

```
ggplot(data=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.

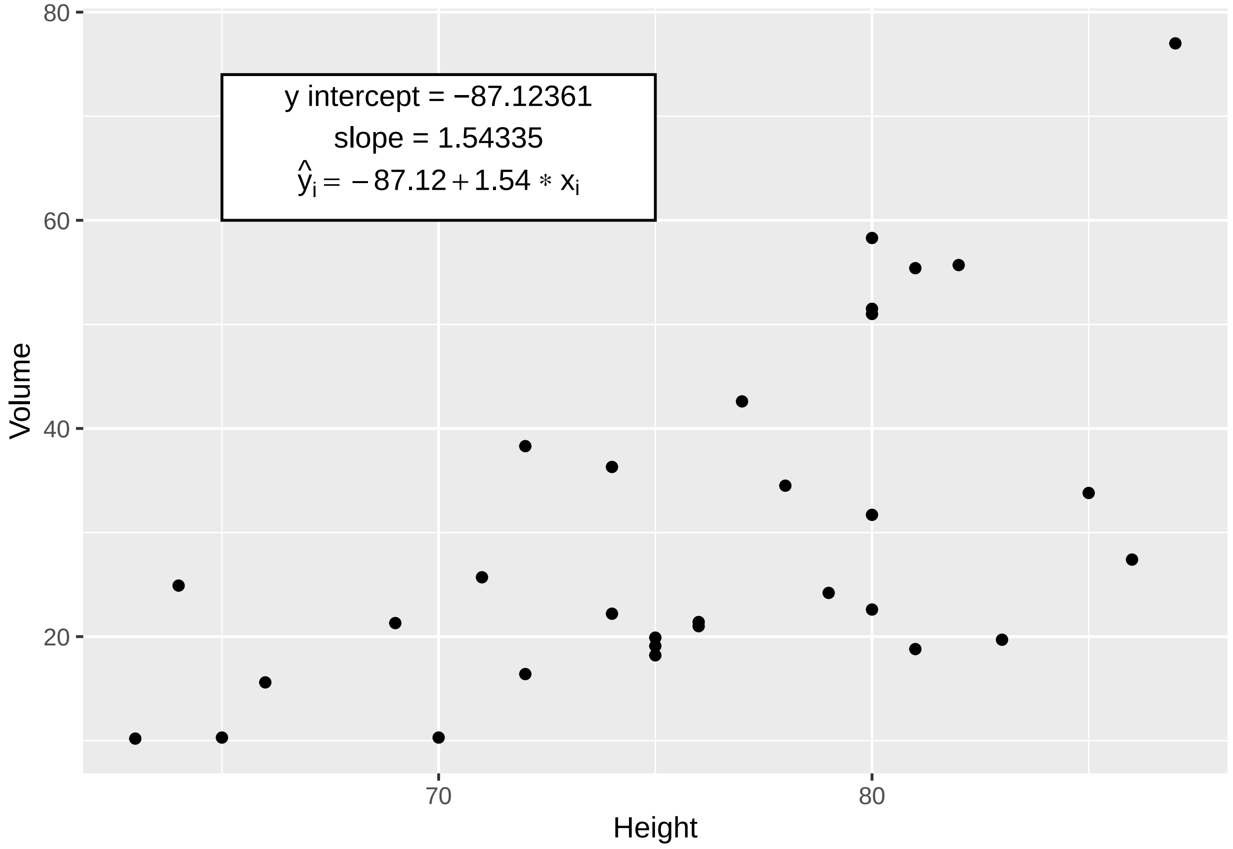
```
ggplot(data=trees, aes(x=Height, y=Volume))+
  geom_point()+
  annotate('rect', xmin=65, xmax=75, ymin=60, ymax=74,
         fill='white', color='black')+
  annotate('text', x=70, y=70,
         label= "y intercept = -87.12361")+
  annotate('text', x=70, y=65,
         label= "slope = 1.54335")
```



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.

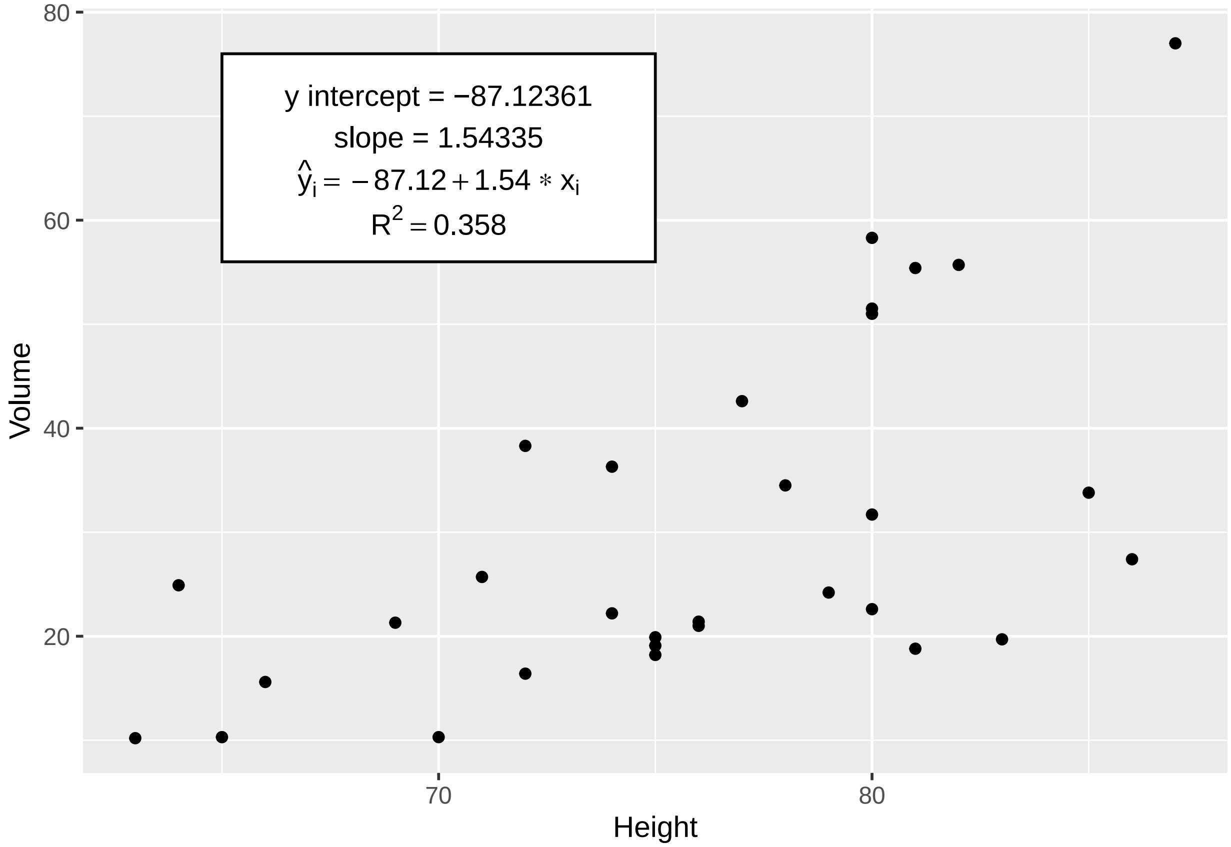
```
ggplot(data=trees, aes(x=Height, y=Volume))+
  geom_point()+
  annotate('rect', xmin=65, xmax=75, ymin=60, ymax=74,
         fill='white', color='black')+
  annotate('text', x=70, y=72,
         label= "y intercept = -87.12361")+
  annotate('text', x=70, y=68,
         label= "slope = 1.54335")+
  annotate('text', x=70, y=64,
         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'
```



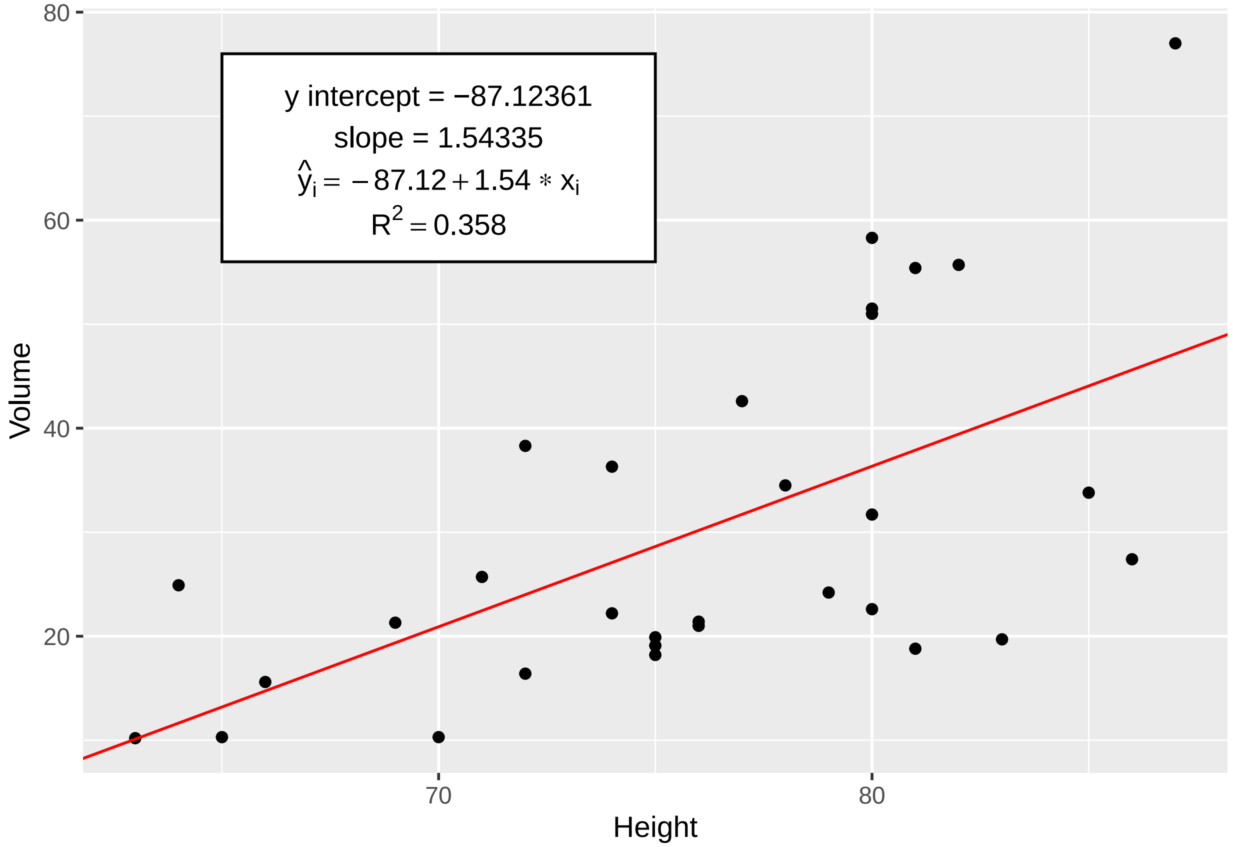
f) Add annotation to add $R^2 = 0.358$

```
ggplot(data=trees, aes(x=Height, y=Volume))+
  geom_point()+
  annotate('rect', xmin=65, xmax=75, ymin=56, ymax=76,
         fill='white', color='black')+
  annotate('text', x=70, y=72,
         label= "y intercept = -87.12361")+
  annotate('text', x=70, y=68,
         label= "slope = 1.54335")+
  annotate('text', x=70, y=64,
         label= latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$'))+
  annotate('text', x=70, y=60,
         label= latex2exp::TeX('$R^2 = 0.358$'))
```



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

```
ggplot(data=trees, aes(x=Height, y=Volume))+
  geom_point()+
  annotate('rect', xmin=65, xmax=75, ymin=56, ymax=76,
         fill='white', color='black')+
  annotate('text', x=70, y=72,
         label= "y intercept = -87.12361")+
  annotate('text', x=70, y=68,
         label= "slope = 1.54335")+
  annotate('text', x=70, y=64,
         label= latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$'))+
  annotate('text', x=70, y=60,
         label= latex2exp::TeX('$R^2 = 0.358$'))+
  geom_abline(slope=1.54335, intercept= -87.12361, color="red")
```

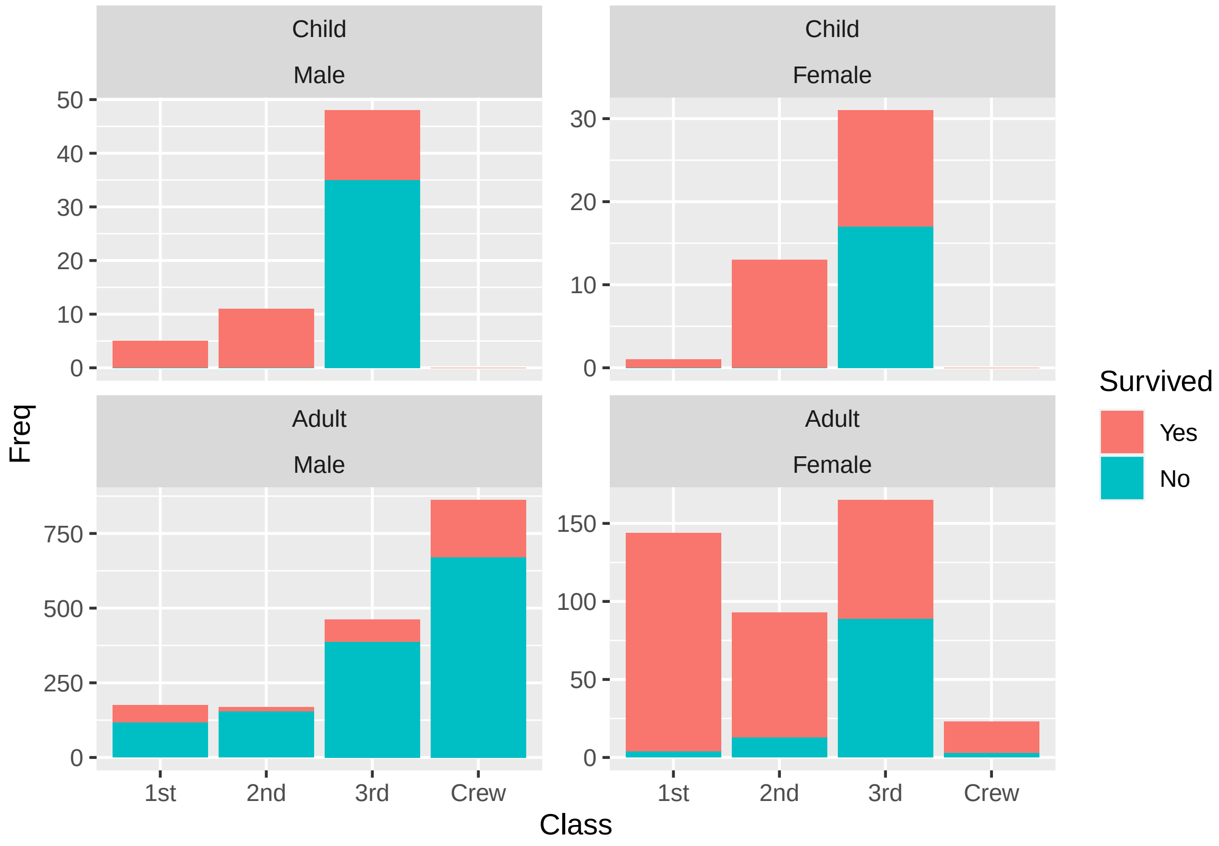


Question 3

In datasets::Titanic table summarizes the survival of passengers aboard the ocean liner Titanic. It includes information about passenger class, sex, and age (adult or child). Create a bar graph showing the number of individuals that survived based on the passenger Class, Sex, and Age variable information. You'll need to use faceting and/or color to get all four variables on the same graph. Make sure that differences in survival among different classes of children are perceivable. Unfortunately, the data is stored as a table and to expand it to a data frame, the following code can be used.

a. Make this graph using the default theme. If you use color to denote survivorship, modify the color scheme so that a cold color denotes death.

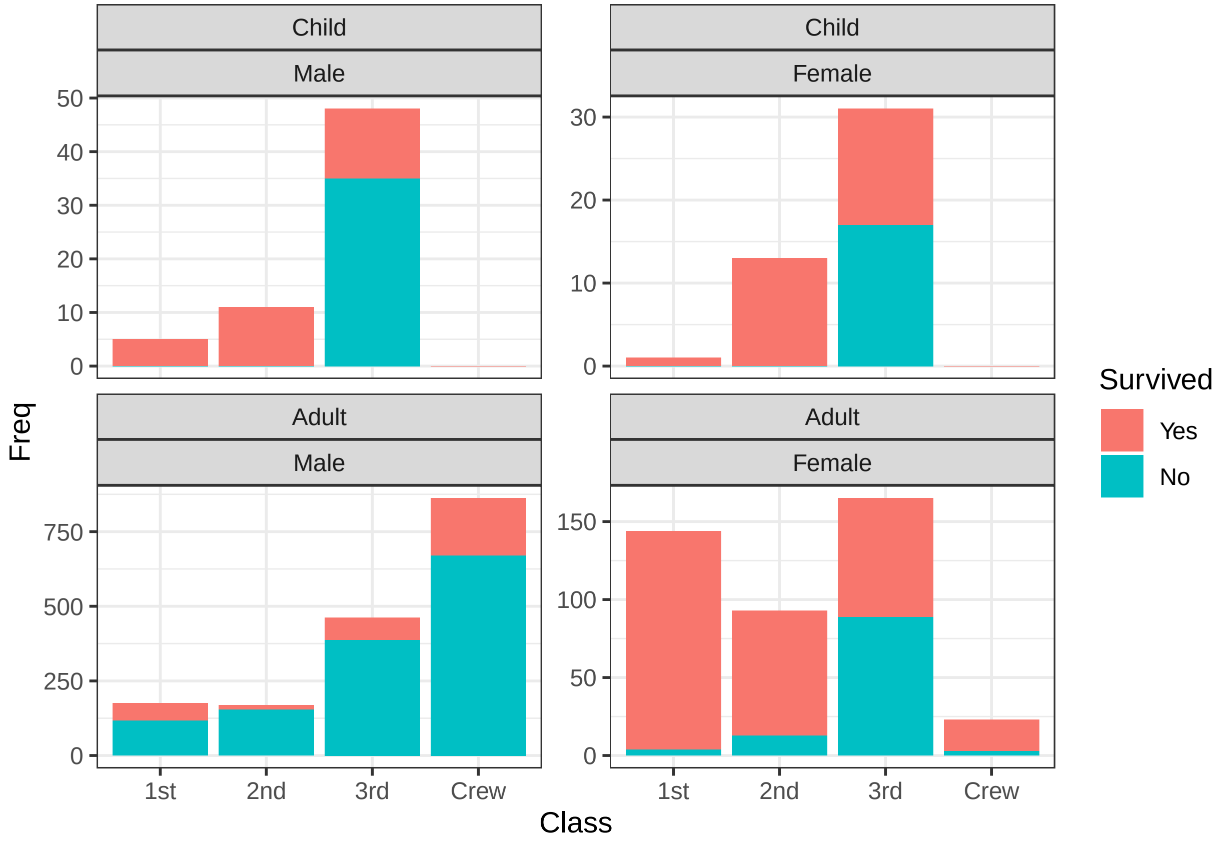
```
Titanic <- Titanic %>% as.data.frame()%>%
  mutate(Survived=fct_relevel(Survived, "Yes", "No"))

ggplot(data=Titanic)+
  geom_col(aes(x=Class, y=Freq, fill=Survived))+
  facet_wrap(Age~Sex, scales="free_y")
```



```
#frequency range is different for each facet
```

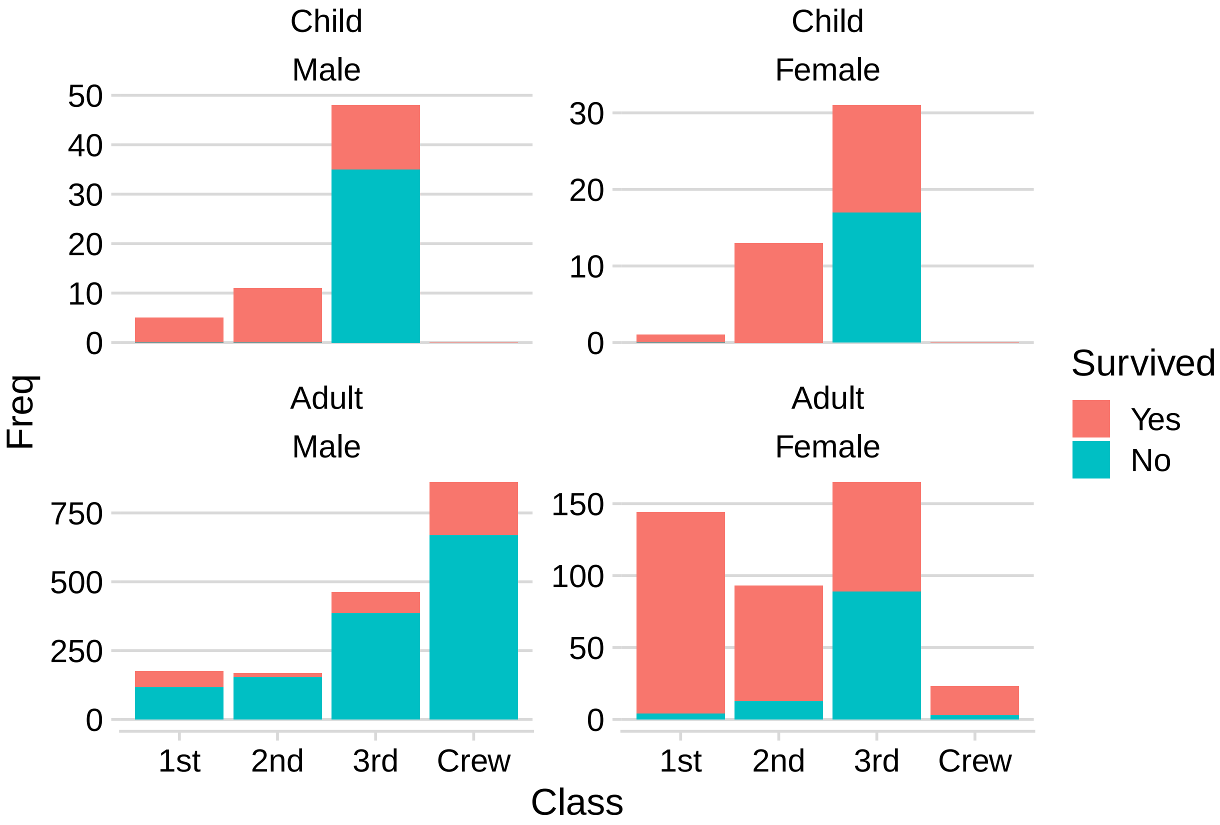
b. Make this graph using the theme_bw() theme.

```
ggplot(data=Titanic)+
  geom_col(aes(x=Class, y=Freq, fill=Survived))+
  facet_wrap(Age~Sex, scales="free_y")+
  theme_bw()
```



c. Make this graph using the cowplot::theme_minimal_hgrid() theme.

```r
library(cowplot)
ggplot(data=Titanic)+
  geom_col(aes(x=Class, y=Freq, fill=Survived))+
  facet_wrap(Age~Sex, scales="free_y")+
  cowplot::theme_minimal_hgrid()
```



d. Why would it be beneficial to drop the vertical grid lines?

Class is a category, so the vertical grid lines are not necessary as each column represents a distinct group.