STA_445_Assignment 7

April Meadows

2024-04-04

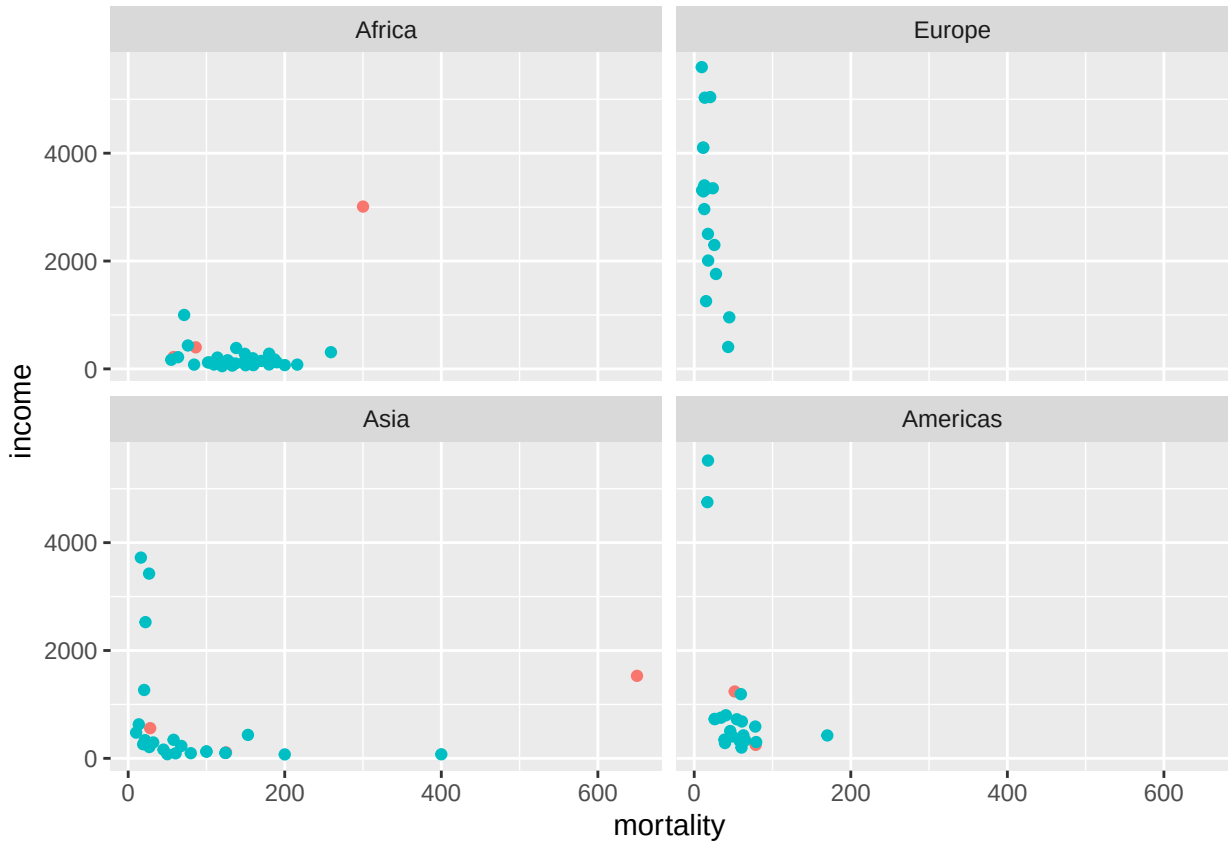
Load your packages here:

Problem 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 The information is relatively out of date, but will be fun to graph.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.

```
# loads the infmort data in
data("infmort")

# plots mortality vs income faceted by region and colored by oil
ggplot(data=infmort,aes(x=mortality,y=income, color=oil)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(infmort$region)
```

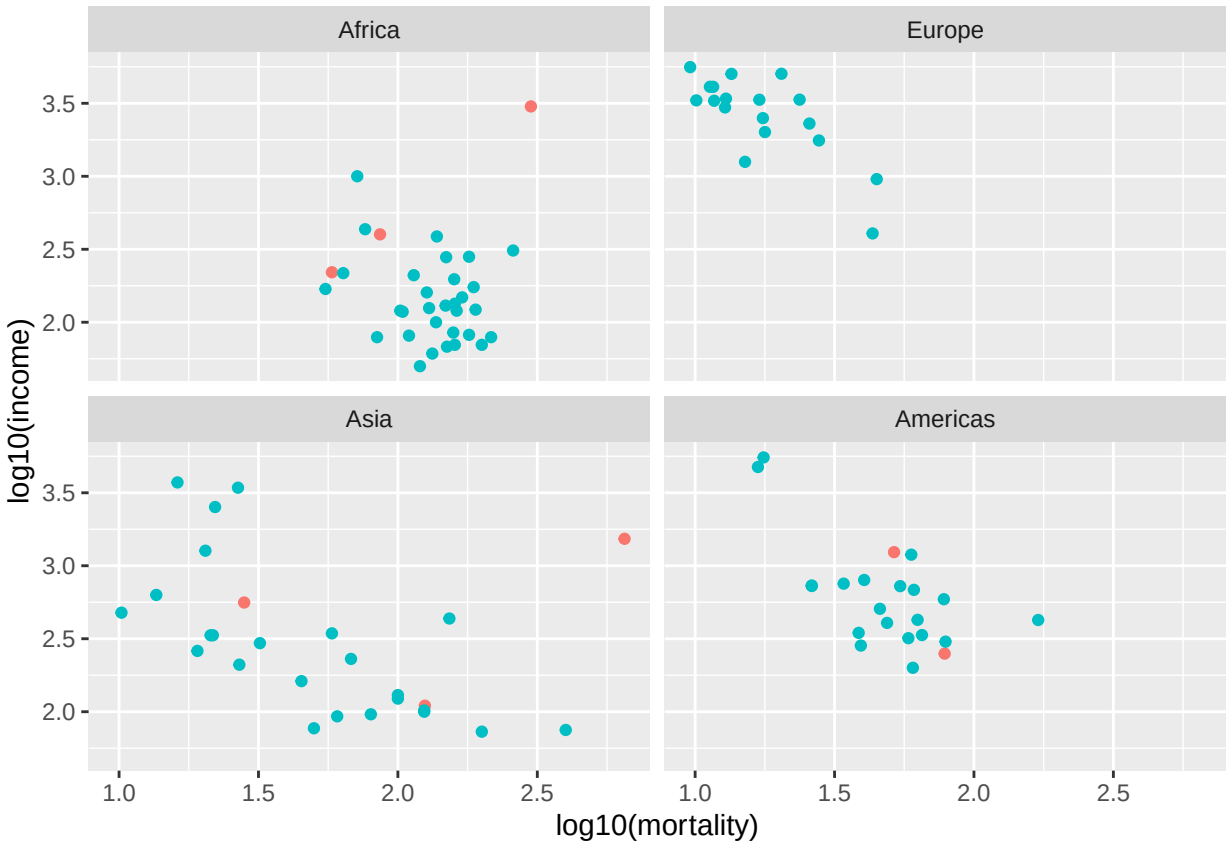


a. The `rownames()` of the table gives the country names and you should create a new column that contains the country names. *`rownames`

```
# adds a column called rownames with the name of each country
infmort <- infmort %>%
  mutate(rownames=rownames(infmort))
```

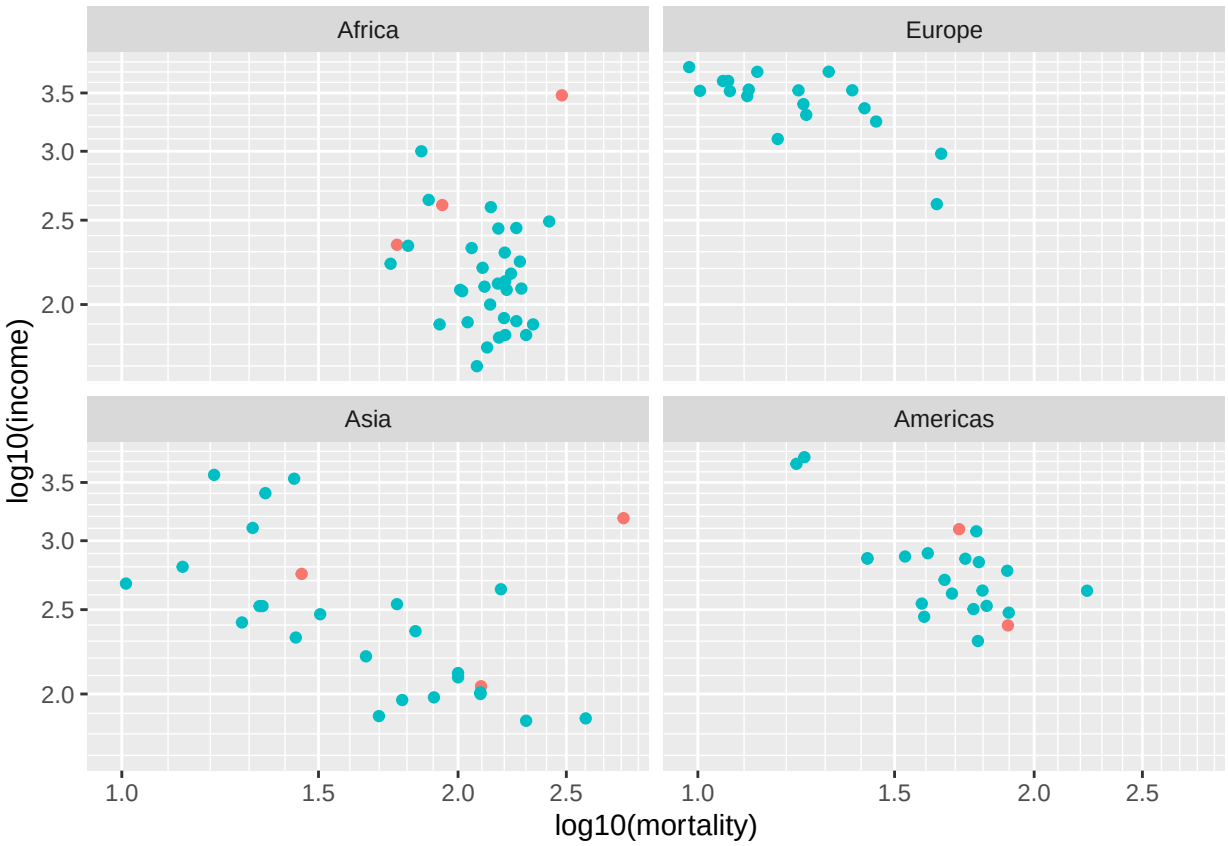
b. Create scatter plots with the `log10()` transformation inside the `aes()`command.

```
# plots mortality vs income with a log10 transformation faceted by region and colored by oil
ggplot(data=infmort,aes(x=log10(mortality),y=log10(income), color=oil)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(infmort$region)
```



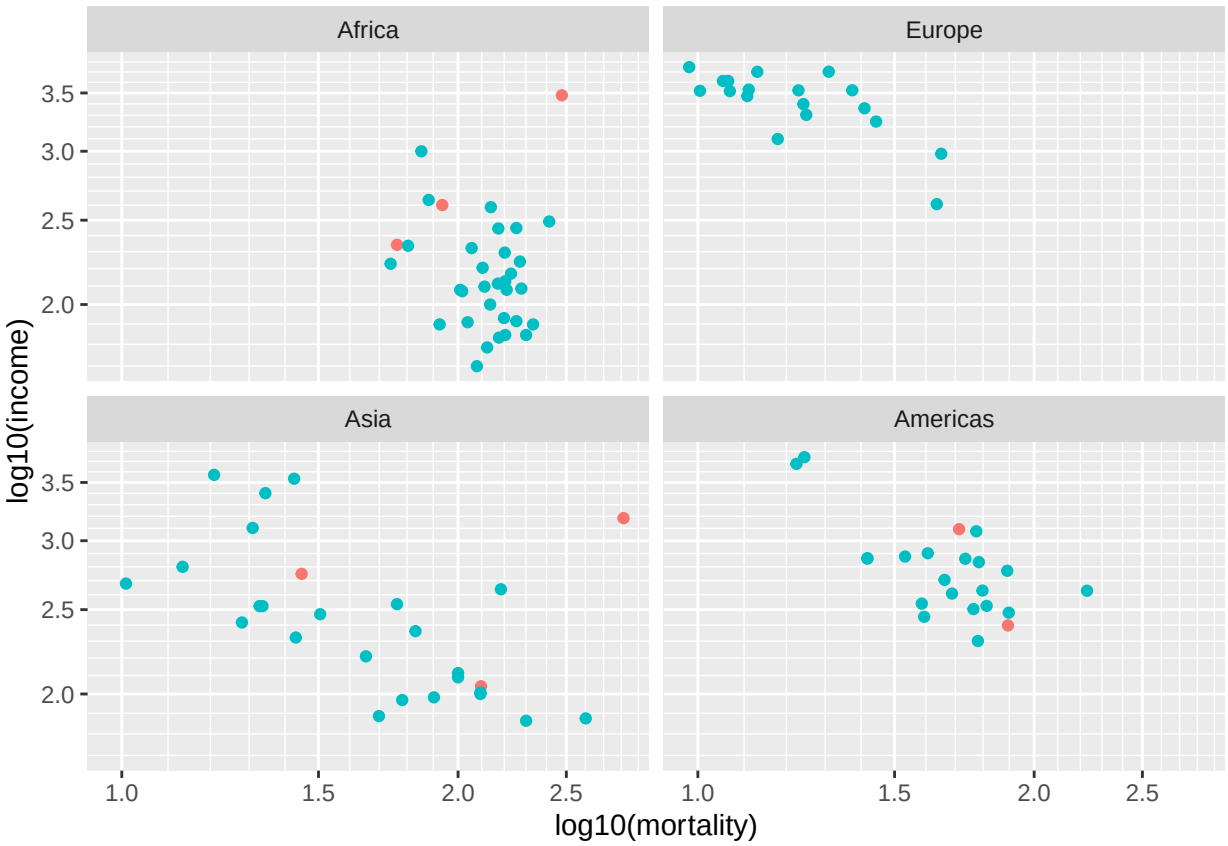
c. Create the scatter plots using the `scale_x_log10()` and `scale_y_log10()`.Set the major and minor breaks to be useful and aesthetically pleasing.Comment on which version you find easier to read.

```
# plots mortality vs income with a log10 transformation faceted by region and colored by oil
ggplot(data=infmort,aes(x=log10(mortality),y=log10(income), color=oil)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(infmort$region) +
  scale_x_log10(breaks=c(1,1.5,2,2.5),
                minor=seq(0,3,0.1)) +
  scale_y_log10(breaks=c(1.5,2,2.5,3,3.5),
                minor=seq(1.5,4,0.1))
```



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.

```
# plots mortality vs income with a log10 transformation faceted by region and colored by oil
ggplot(data=infmort,aes(x=log10(mortality),y=log10(income), color=oil,label=infmort$rownames)) +
  geom_point(show.legend = FALSE) +
  facet_wrap(infmort$region) +
  scale_x_log10(breaks=c(1,1.5,2,2.5),
                minor=seq(0,3,0.1)) +
  scale_y_log10(breaks=c(1.5,2,2.5,3,3.5),
                minor=seq(1.5,4,0.1)) +
  # geom_text_repel(force=100,size=3) +
  theme(legend.position = "none")
```



geomrepel was not working for me. The document would not knit with it due to the unlabeled data points. I created a sub-data frame with 10 countries in it, but I kept getting an error since the two data frames were different sizes. I spent over an hour on this one section of the assignment so I gave up here :(

Problem 2

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

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

```
# loads trees data in
data("trees")

# saves linear model of Height vs Volume in Trees
Trees <- lm(Volume~Height, data=trees)
```

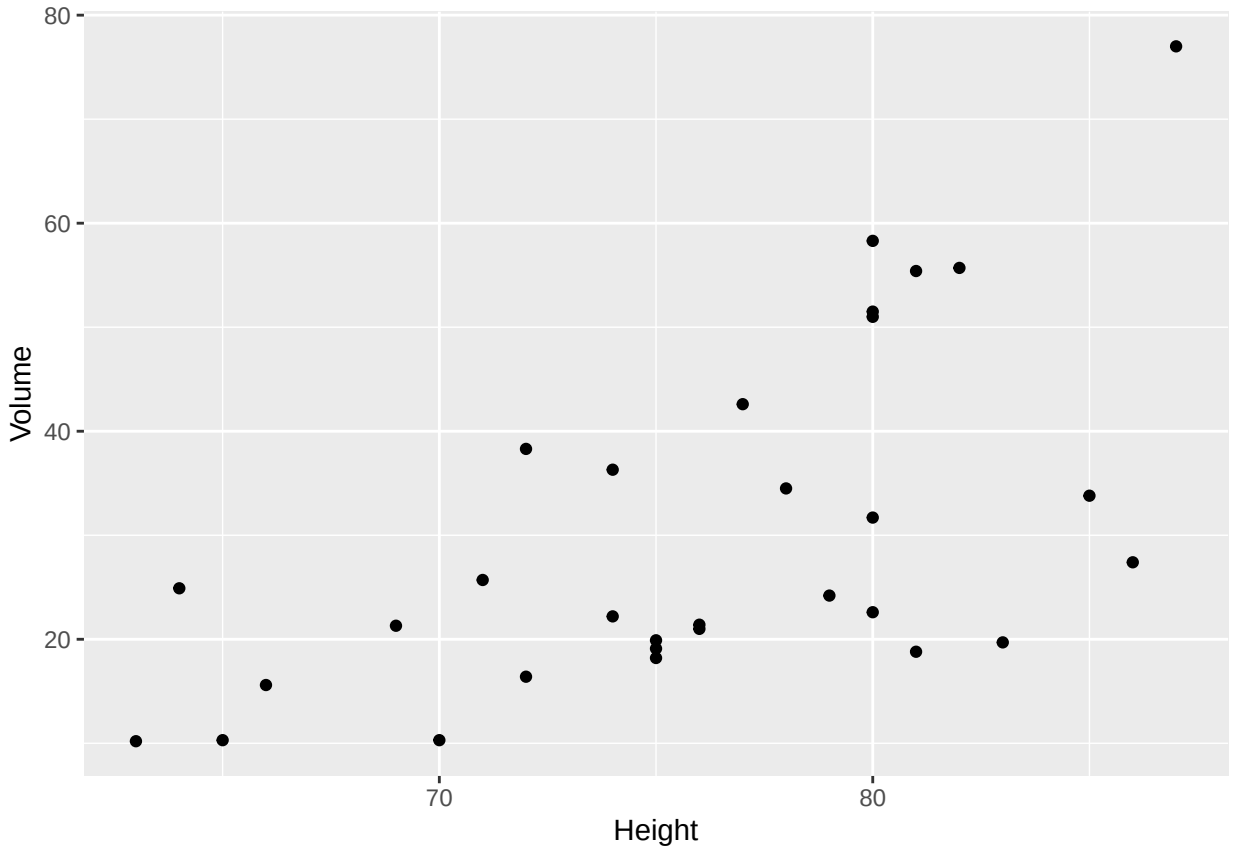
b. Using the str(your model's name) command, to get a list of all the information stored in the linear model object. Use $ to extract the slope and intercept of the regression line (the coefficients).

```
# shows the coefficients of lm Trees
str(Trees$coefficients)
```

```
##  Named num [1:2] -87.12 1.54
##  - attr(*, "names")= chr [1:2] "(Intercept)" "Height"
```

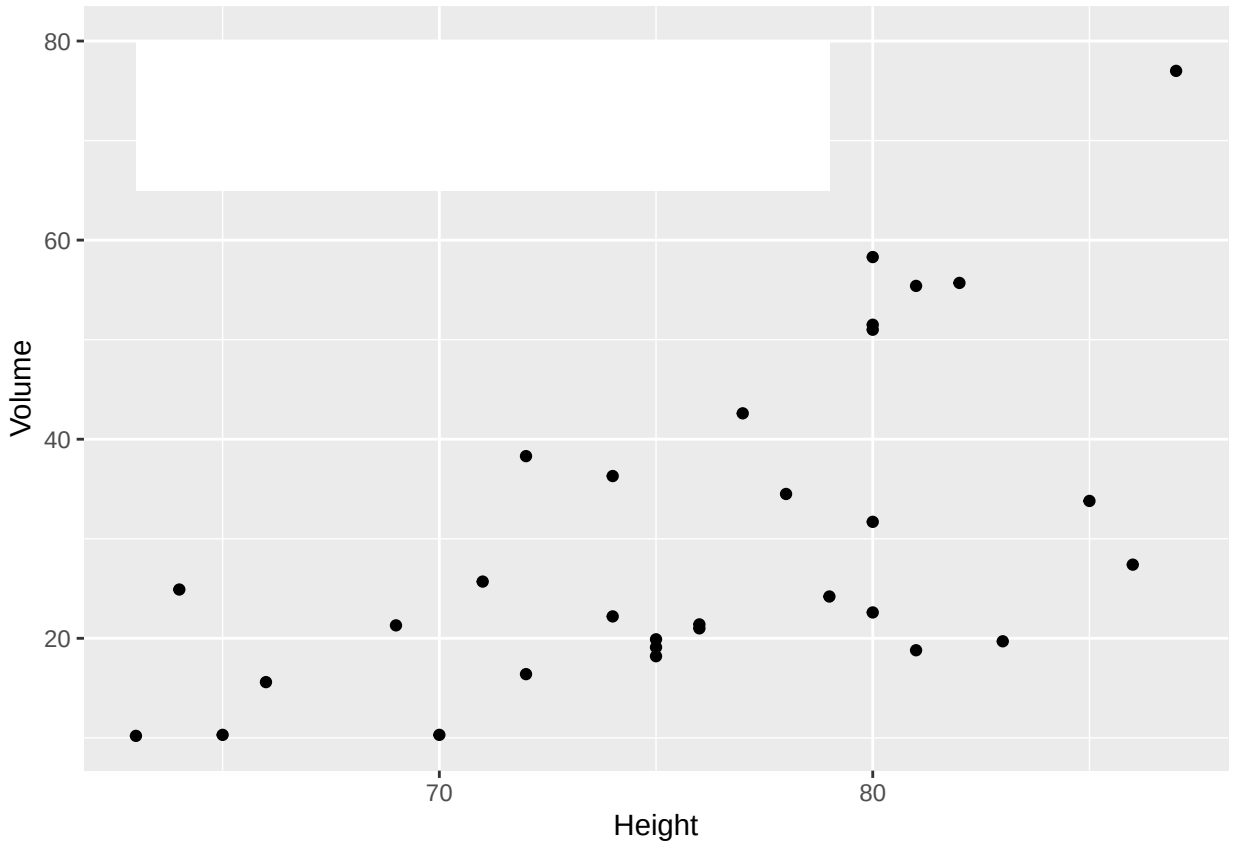
c. Using `ggplot2`, create a scatter plot of Volume vs Height.

```
# plots Height vs Volume
ggplot(data=trees, aes(x=Height,y=Volume)) +
  geom_point()
```



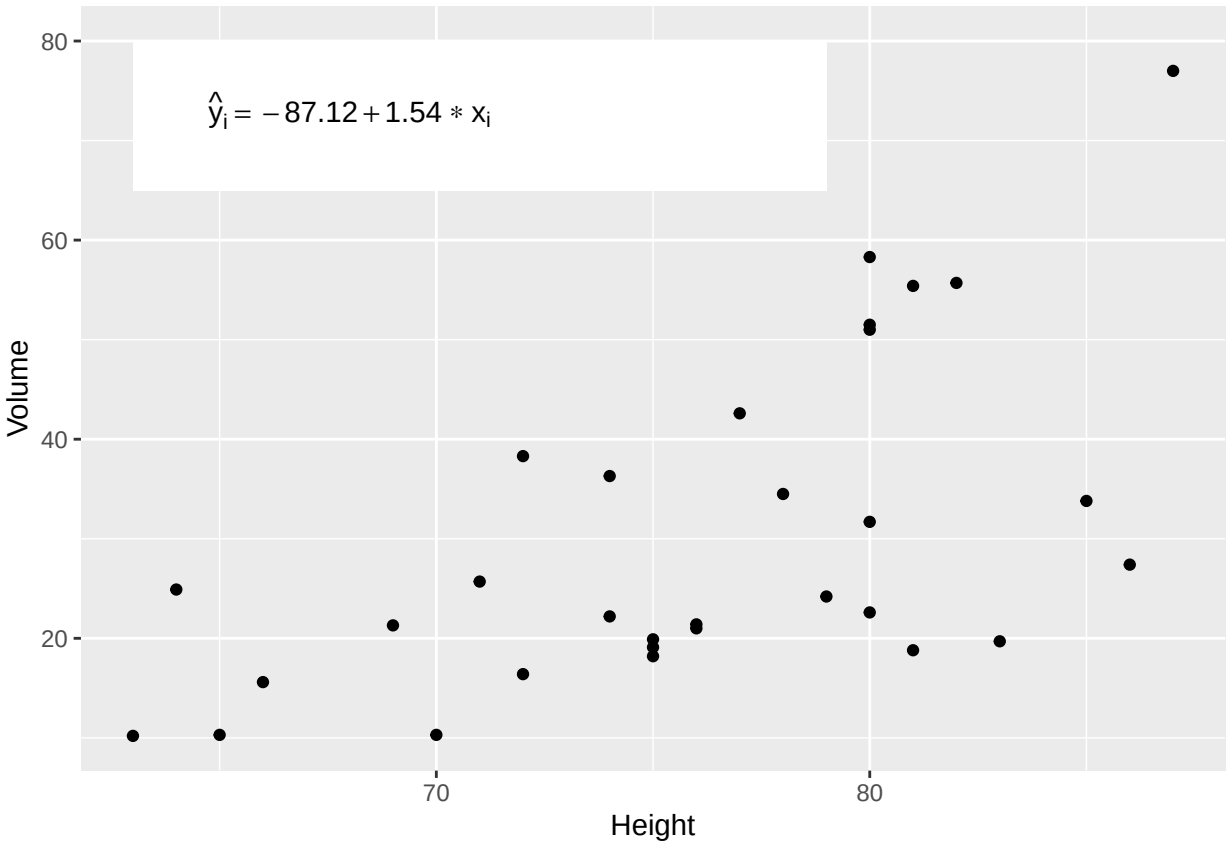
d. Create a nice white filled rectangle to add text information to using by adding the following annotation layer.

```
# plots Height vs Volume and adds a white rectangle at the top left
ggplot(data=trees, aes(x=Height,y=Volume)) +
  geom_point() +
  annotate('rect',xmin=63,ymin=65,xmax=79,ymax=80,
           fill="white")
```



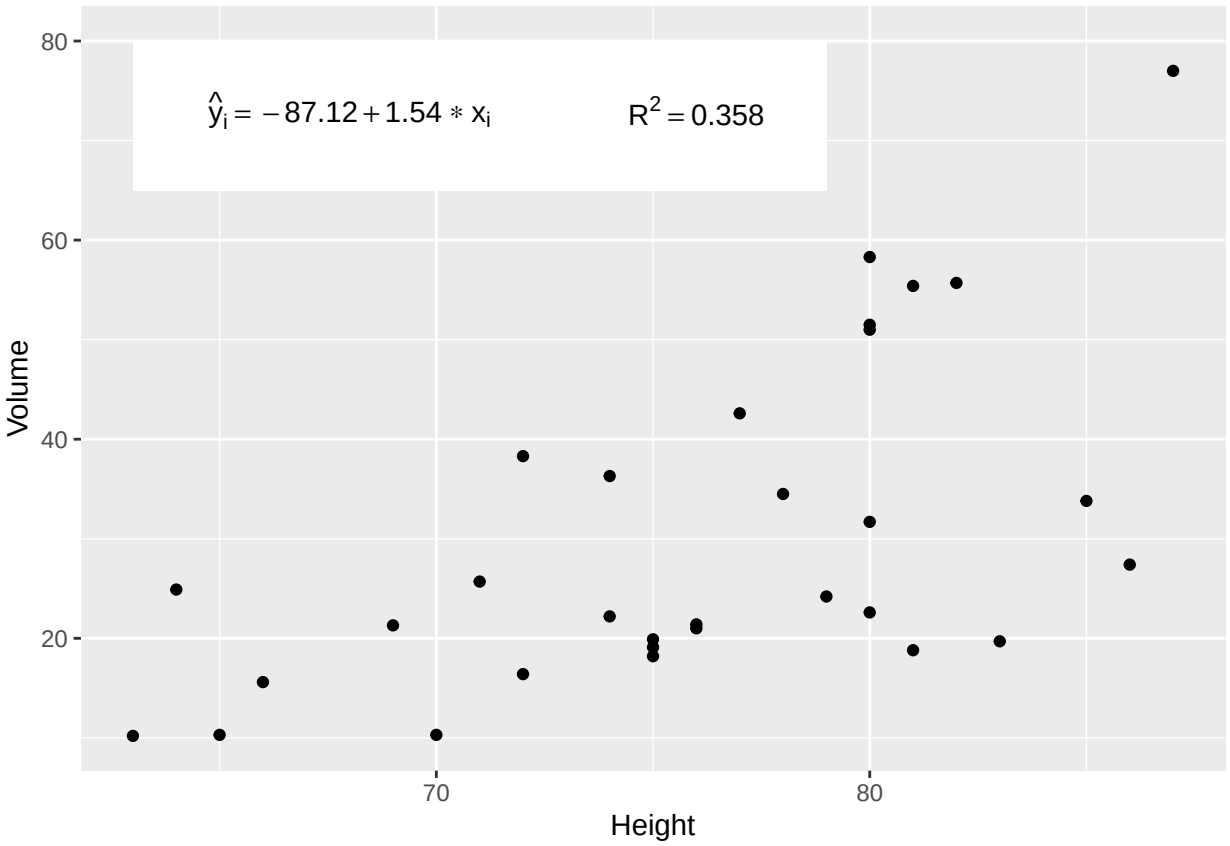
e. Add some annotation text to write the equation of the line $\hat{y}_i = -87.12 + 1.54 * x_i$ in the text area.

```
# plots Height vs Volume with a white rectangle at the top left and adds the linear model
# equation in it
ggplot(data=trees, aes(x=Height,y=Volume)) +
  geom_point()+
  annotate('rect',xmin=63,ymin=65,xmax=79,ymax=80,
           fill="white") +
  annotate('text',x=68,y=73,
           label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$'))
```



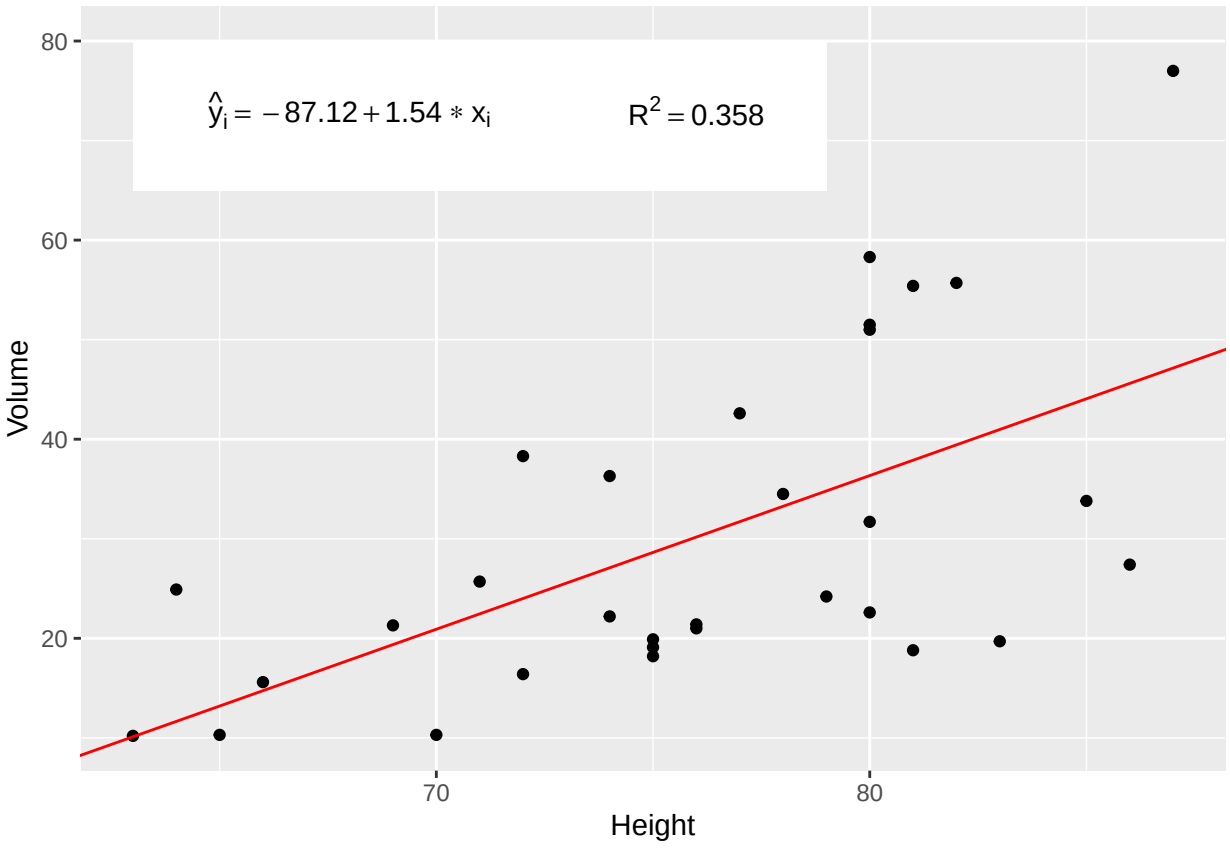
f. Add annotation to add $R^2 = 0.358$

```
# plots Height vs Volume with a white rectangle at the top left with the linear model
# equation and adds r^2 in it
ggplot(data=trees, aes(x=Height,y=Volume)) +
  geom_point()+
  annotate('rect',xmin=63,ymin=65,xmax=79,ymax=80,
           fill="white") +
  annotate('text',x=68,y=73,
           label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$')) +
  annotate('text',x=76,y=73,label=latex2exp::TeX('$R^2 = 0.358$'))
```



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

```
# plots Height vs Volume with a white rectangle at the top left with the linear model
# equation and r^2 in it and adds a red lm line
ggplot(data=trees, aes(x=Height,y=Volume)) +
  geom_point() +
  annotate('rect',xmin=63,ymin=65,xmax=79,ymax=80,
           fill="white") +
  annotate('text',x=68,y=73,
           label=latex2exp::TeX('$\\hat{y}_i = -87.12 + 1.54 * x_i$')) +
  annotate('text',x=76,y=73,label=latex2exp::TeX('$R^2 = 0.358$')) +
  geom_abline(slope=Trees$coefficients[2],intercept =Trees$coefficients[1],col="red")
```



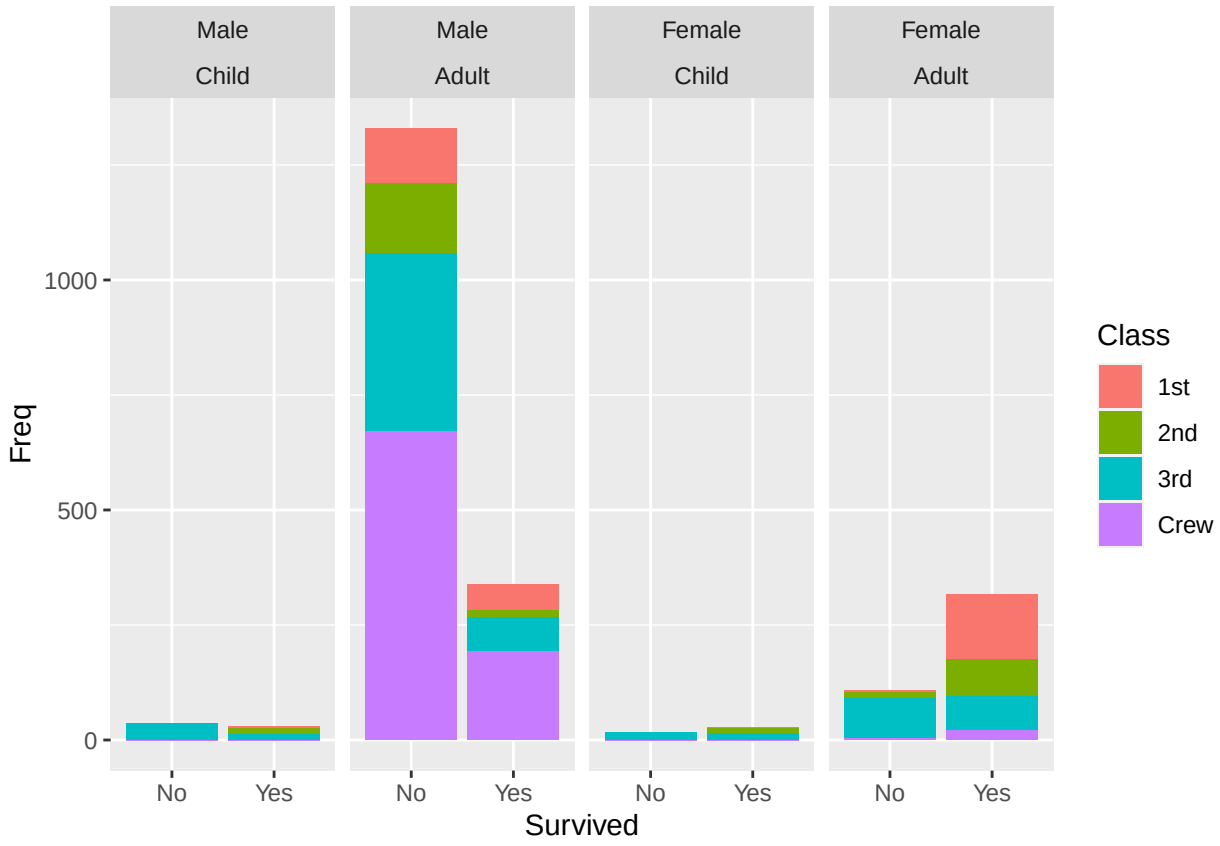
Problem 3

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.*

```r
Titanic <- Titanic %>% as.data.frame()
```

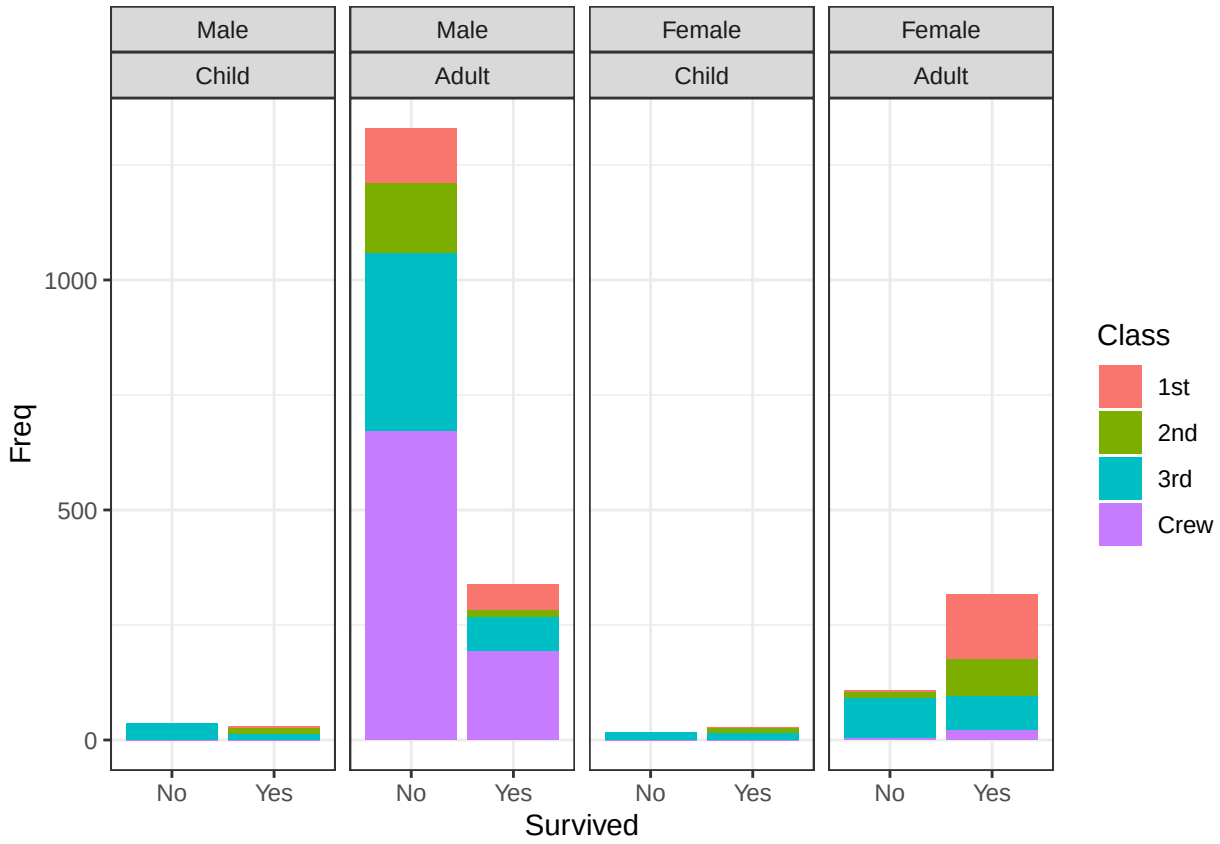
a. Make this graph using the default theme. *If you use color to denote survivorship, modify the color scheme so that a cold color denotes death.*

```r
# plots a bar graph of survival status vs frequency
# graph is colored by class and faceted by sex and age
ggplot(data=Titanic, aes(x=Survived,y=Freq,fill=Class)) +
  geom_bar(stat="identity") +
  facet_wrap(facets=c("Sex","Age"),nrow=1)
```



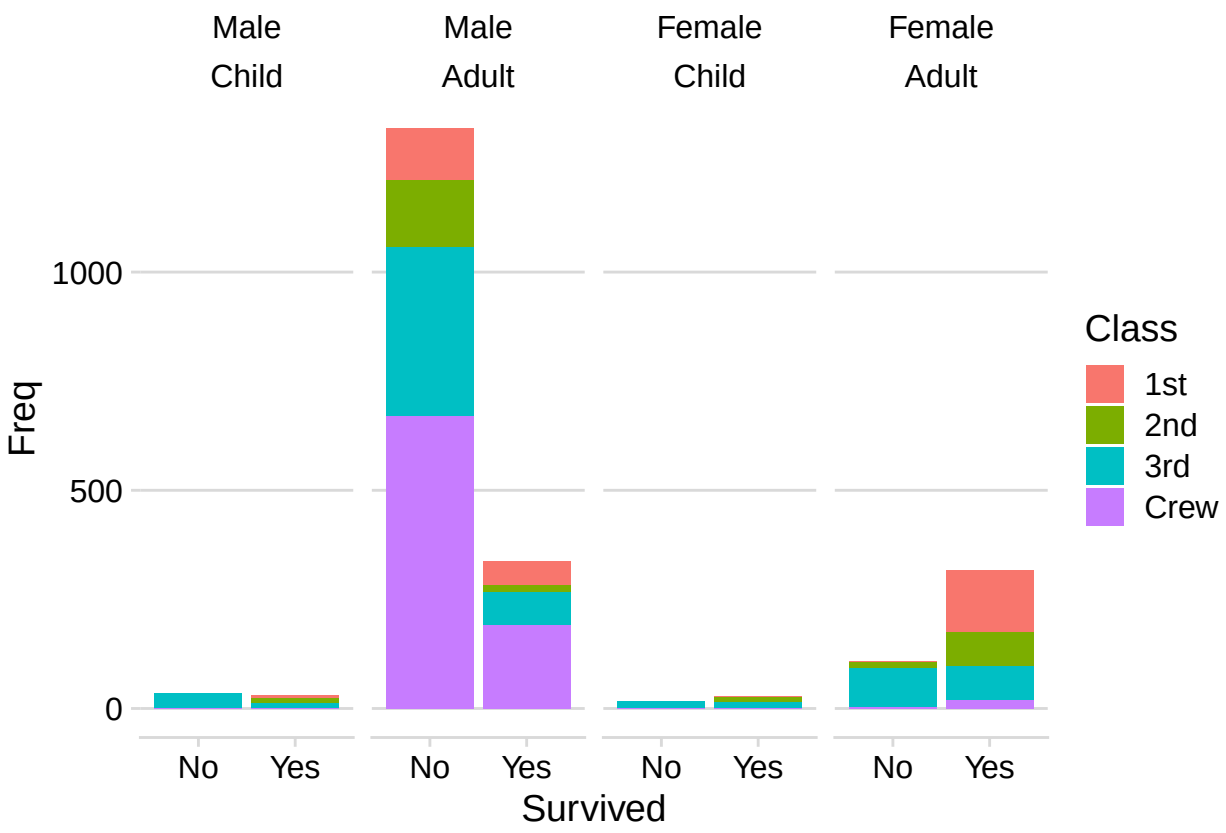
b. Make this graph using the `theme_bw()` theme.

```
# plots a bar graph of survival status vs frequency
# graph is colored by class and faceted by sex and age
# adds the bw theme to it
ggplot(data=Titanic, aes(x=Survived,y=Freq,fill=Class)) +
  geom_bar(stat="identity") +
  facet_wrap(facets=c("Sex","Age"),nrow=1) +
  theme_bw()
```



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

```
# plots a bar graph of survival status vs frequency
# graph is colored by class and faceted by sex and age
# adds a cowplot theme to it
ggplot(data=Titanic, aes(x=Survived,y=Freq,fill=Class)) +
  geom_bar(stat="identity") +
  facet_wrap(facets=c("Sex","Age"),nrow=1) +
  cowplot::theme_minimal_hgrid()
```



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

Vertical grid lines typically help indicate what the values along the x-axis are. Our x-axis does not have a scale of values, so the vertical lines are not necessary and may even cause some confusion.