Evan Krook

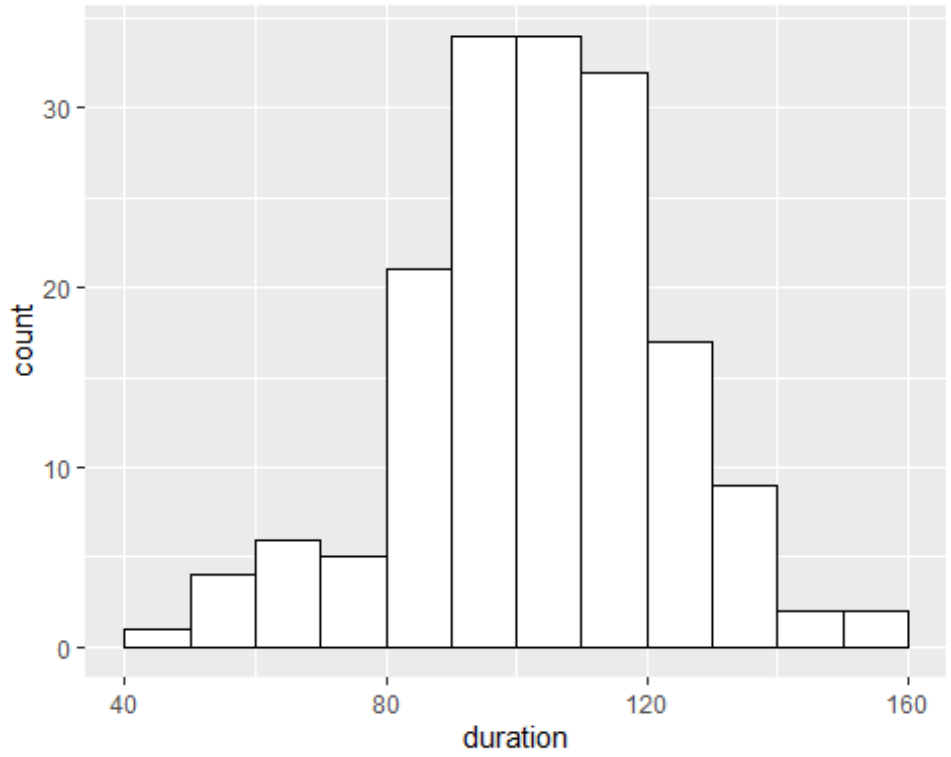
Working with Graphical Visualization of Data

Problems

1

The following code makes a histogram of the `duration` variable in the Lake Monona data set.

```
ggplot(monona, aes(x=duration)) +
  geom_histogram(boundary = 0, binwidth = 10,
                 color = "black", fill = "white")
```



In approximately how many winters was the total duration where Lake Monona was at least 50% covered with ice between 40 to 70 days?

Response

Approximately 11 winter was lake monona 50% covered with ice between 40 to 70 days.

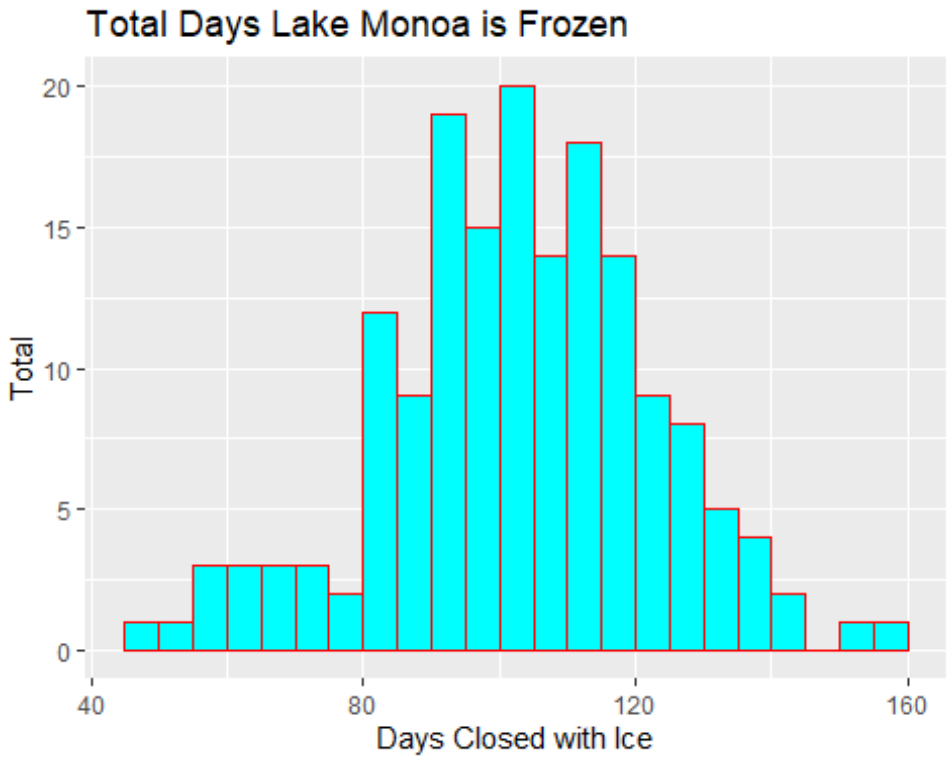
2

Modify the code below so that:

- one of the bin boundaries is at 70 days

- the width of each bin is 5 days
- the fill color is “cyan”
- the color outlining the bars is “red”
- the x label says “Days Closed with Ice”
- the y label says “Total”
- there is a title with words of your choosing that describe the figure

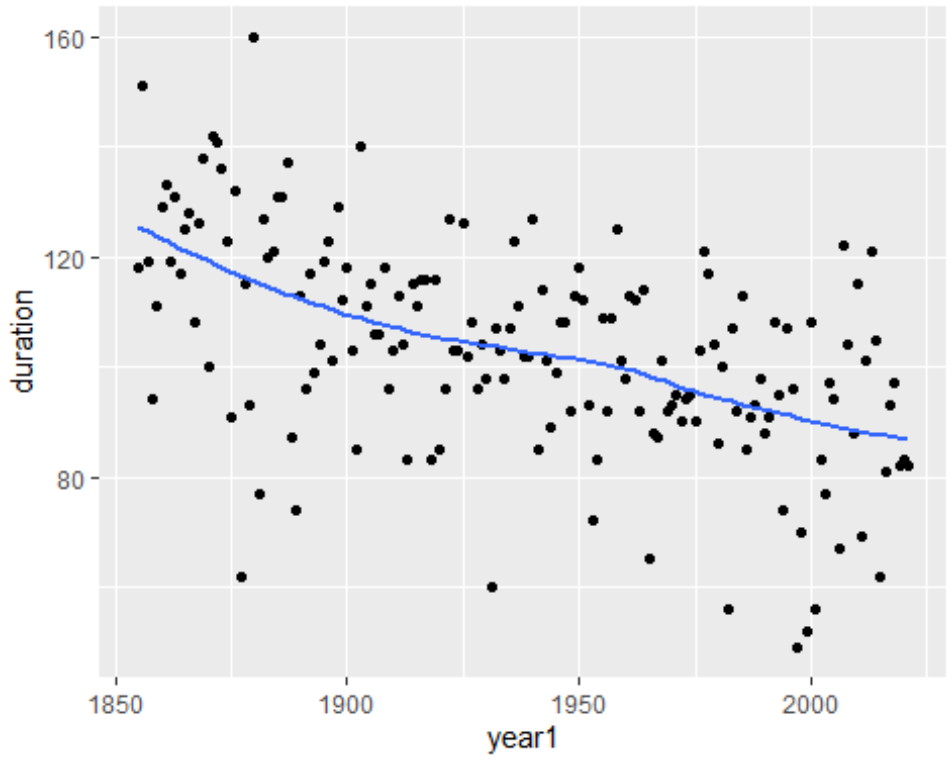
```
ggplot(monona, aes(x = duration)) +
  geom_histogram(boundary = 70, binwidth = 5, fill = "cyan", color = "red") +
  xlab("Days Closed with Ice") +
  ylab("Total") +
  ggtitle("Total Days Lake Monoa is Frozen")
```



3

Code in the next chunk makes a scatter plot that shows how the variable `duration` changes with time (using `year1`).

```
ggplot(monona, aes(x = year1, y = duration)) +
  geom_point() +
  geom_smooth(se=FALSE)
```



- What does the line of code `geom_smooth(se=FALSE)` do? (Explain what it does on the graphic; you don't need to explain details of the method.)

Response

geom_smooth(se=False) adds a trendline (average) to the scatter plot representing the relationship of all the data present between 'year1' and 'duration' from the Lake Monona data, and the "False" gets rid of the ribbon associated with the trendline.

Response

N/A

- How long was Lake Monona closed with closed with ice in a typical year near 1875 (i.e., what is the approximate value of the smooth curve around 1875)

Response

Near the year 1875, Lake Monona was closed with ice approximately 118 days in a typical year (Just under 120 days).

- How long was Lake Monona closed with ice in a typical year near 2000 (i.e., what is the approximate value of the smooth curve around 2000)?

Response

Near the year 200, Lake Monona was closed with ice approximately 90 days in a typical year (Between 80 and 100 days).

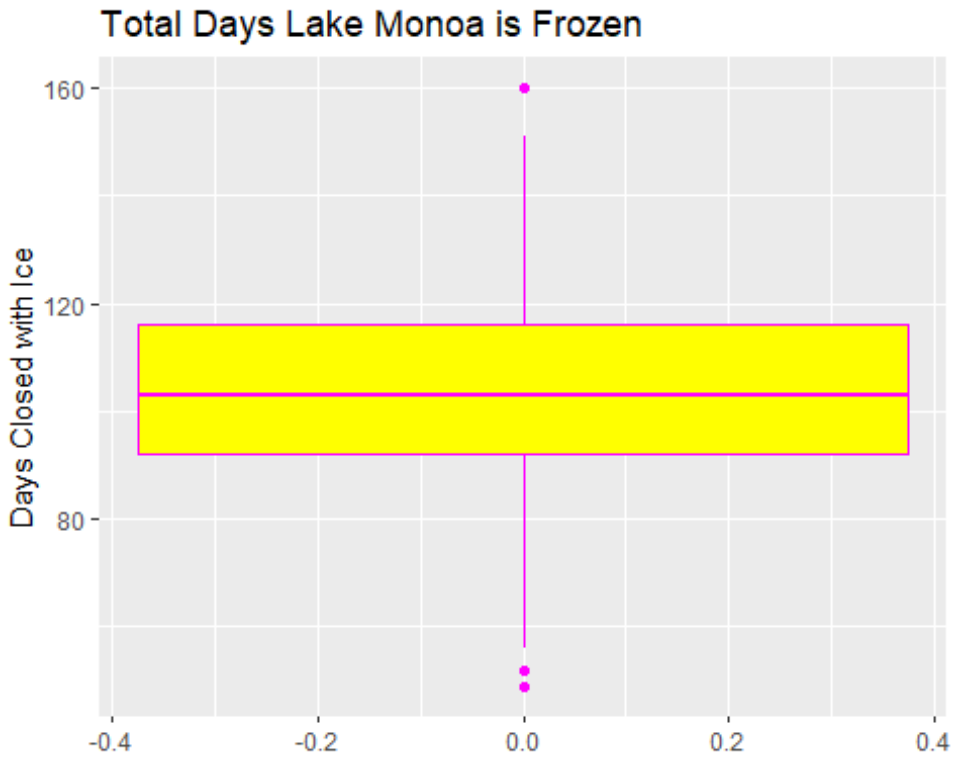
4

Modify the code in the following chunk so that:

- There is a box plot displaying the distribution of the days frozen by ice

- The box plot fill color is “yellow”
- The color of the edges of the box plot is “magenta”
- There is a more descriptive y-axis label
- There is an informative plot title

```
ggplot(monona, aes(y=duration)) +
  geom_boxplot(color = "magenta", fill = "yellow",varwidth = TRUE)+
  ylab("Days Closed with Ice") +
  ggtitle("Total Days Lake Monoa is Frozen")
```



- What is the approximate median number of days Lake Monona has been closed with ice?

Response

Lake Monona median number of days closed with ice is approximately 105 days (Between 100 and 120 days).

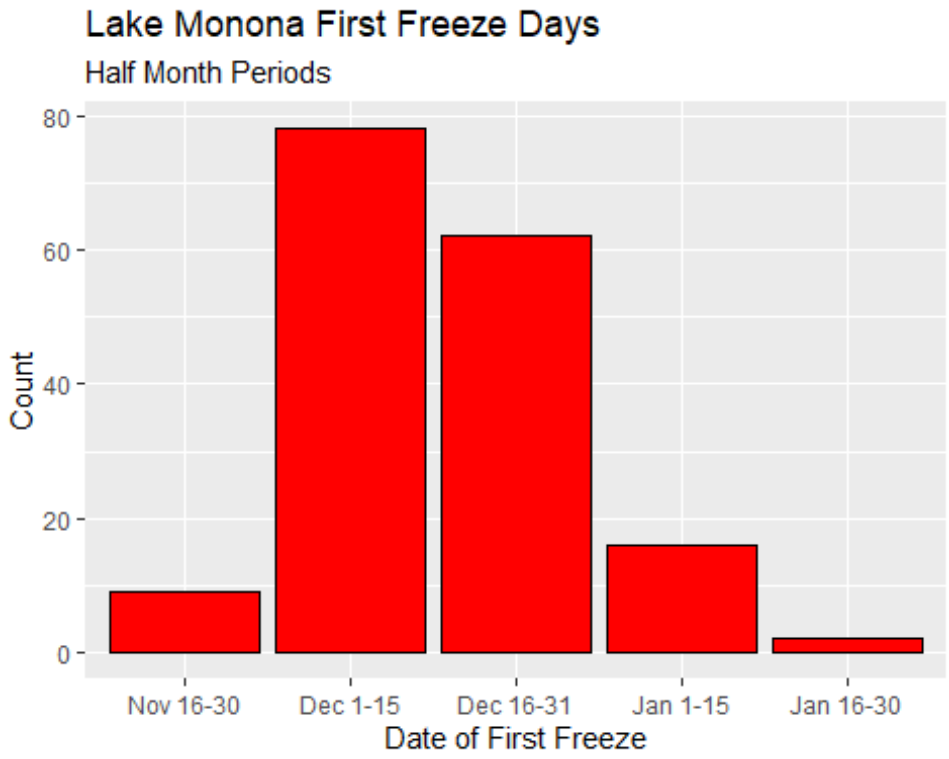
5

- Write code to create a bar graph which displays the number of winters when the first freeze occured in each half-month period of time as recorded in the variable

`ff_cat`. Choose your own colors if you do not like the default values. Make sure that your plot:

- has an informative title and subtitle
- has informative axis labels

```
ggplot(monona, aes(x = ff_cat)) +
  geom_bar(color = "black", fill = "red") +
  xlab("Date of First Freeze") +
  ylab("Count") +
  ggtitle("Lake Monona First Freeze Days", subtitle = "Half Month Periods")
```



6

- Briefly explain why you needed to use the command `geom_bar()` and not `geom_col()` to make the plot in the previous problem.

The use of geom_bar() is needed because we only want the total number of cases of the first freeze happening in each half month period to be represented in the hiehgt, and not having a data representation in the hieght.

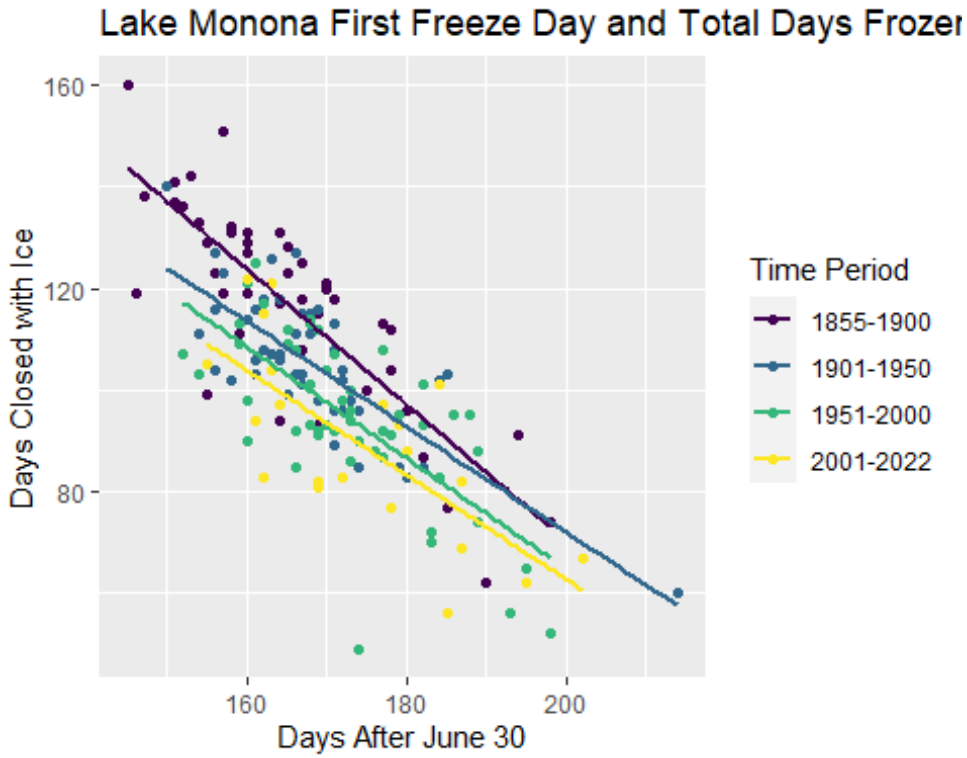
7

- The following chunk creates a scatter plot with `ff_x` on the x axis and `duration` on the y axis, with points colored by `period50`. The variable `ff_x` is a numerical coding of the first freeze date, counting days after June 30. For context, December 27 is 180 days after June 30. The default color scheme is changed to `viridis` which is friendlier to most people with various forms of color blindness. The command `geom_smooth(method = "lm", se = FALSE)` adds a straight line instead of a curve

to the plot (that's the `method = "lm"` argument) and because we specified `period50` as a grouping variable by mapping it to the color aesthetic, separate lines are added for each group.

- Add code to add a plot title and to provide informative axis labels. Following examples from lecture notes, change the title of the color legend to say "Time Period" instead of "period50".

```
ggplot(monona, aes(x = ff_x, y = duration, color = period50)) +
  geom_point() +
  geom_smooth(se = FALSE, method = "lm") +
  scale_color_viridis_d() +
  xlab("Days After June 30") +
  ylab("Days Closed with Ice") +
  labs(colour = "Time Period") +
  ggtitle("Lake Monona First Freeze Day and Total Days Frozen")
```



8

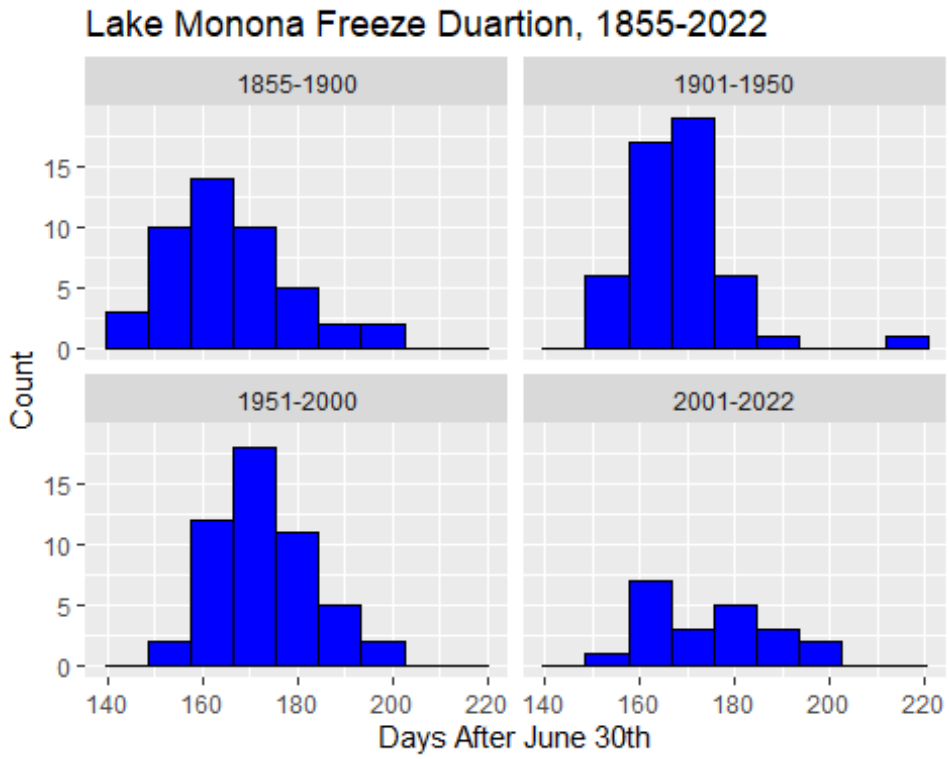
- The graph from the previous problem could be used to predict the total freeze duration of Lake Monona based on the date of the first freeze when the surface of Lake Monona is first at least 50% covered by ice. Suppose that the date of the first freeze in some year was December 27, which is 180 days after June 30. Based on an examination of the graph, briefly explain how your prediction of the total duration that Lake Monona is closed by ice would differ if the winter was in the 1870s versus the present?

My prediction would differ in the 1870s versus the present becasue in the 1870s the trendline is showing approximately 100 days frozen on average with frist freeze coming 180 days after June 30th and the present is showing approximately 75 days frozen on average with first freeze coing 180 days after June 30th. So, I would have a difference of about 25 days Lake Monona is frozen depending on what year I use in research.

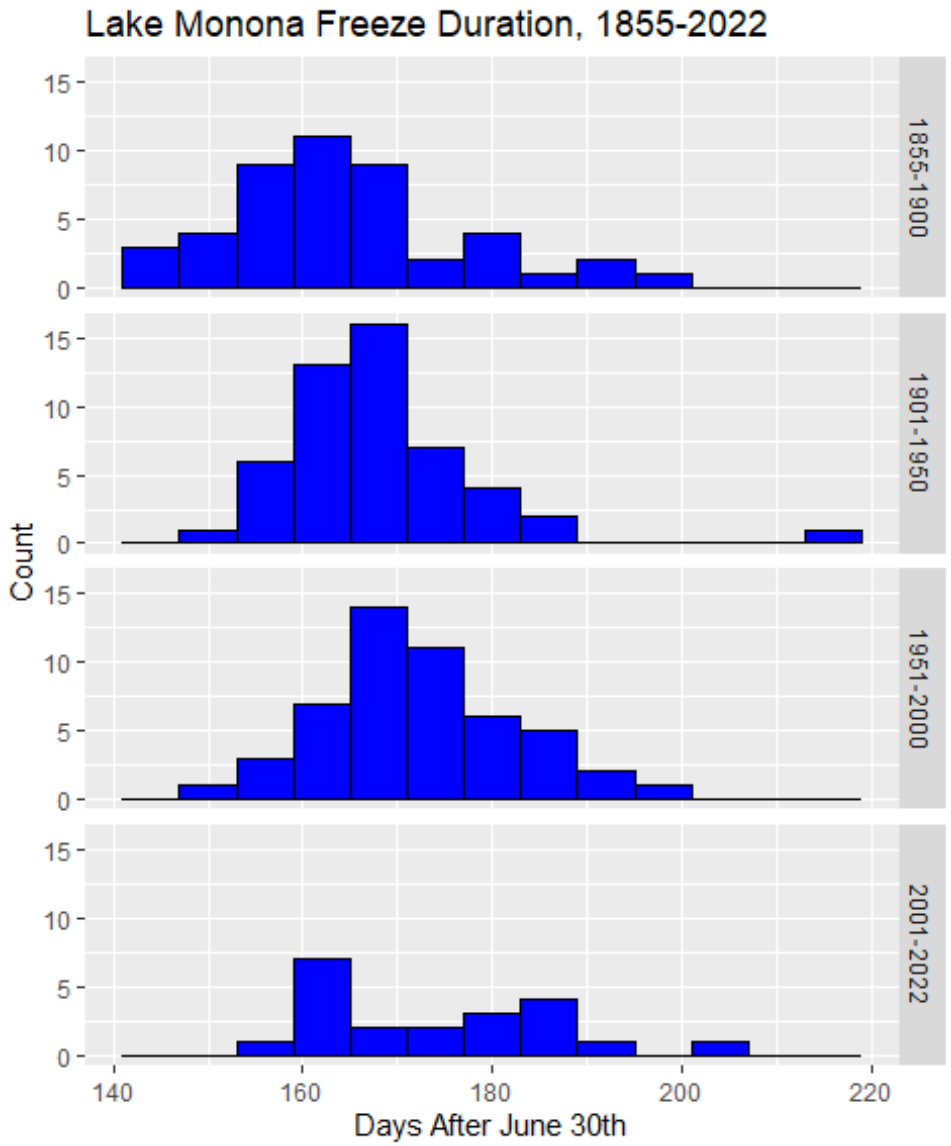
9

- The next four chunks contain partial code to make separate graphs to examine the distribution of the first day of a winter season that Lake Monona is at least 50% closed by ice (coded as a numerical variable which counts days after June 30for each of the four periods time periods encoded in `period50`.
  - Separate histograms "wrapped" to make a linear sequence of plots broken over one or more rows.
  - Separate histograms arranged to be stacked and using the set of axes.
  - Side-by-side box plots.
  - Density plots overlaid on the same graph.
- For each chunk, complete the code to produce the desired graph. Add informative axis labels and graph titles.

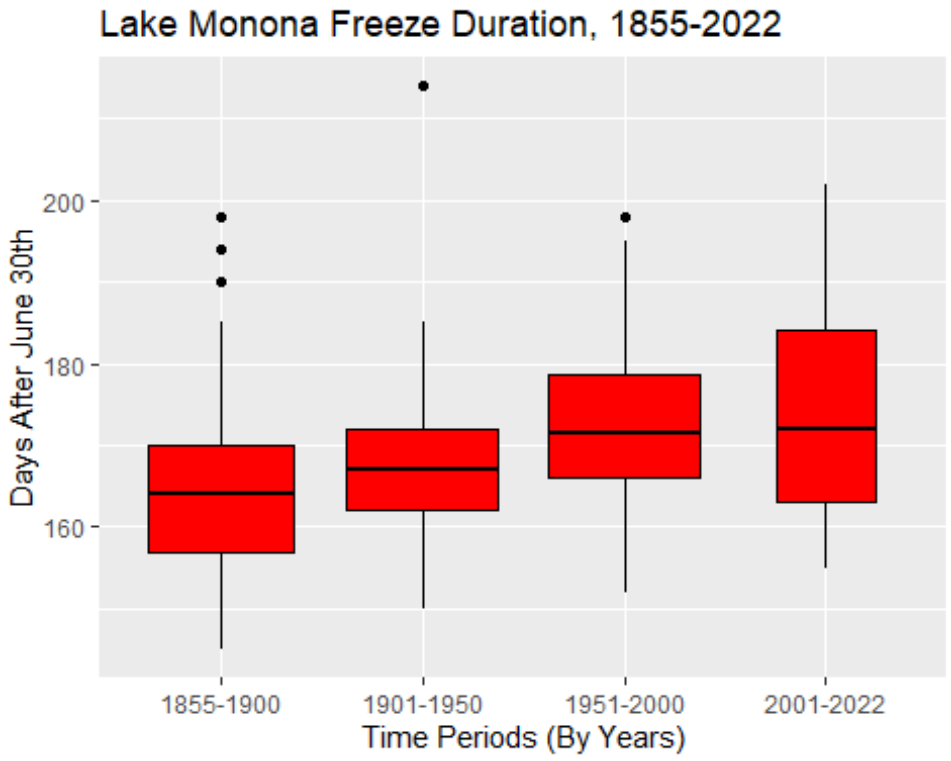
```
## wrapped histograms
##
## change the binwidth to an appropriate value
## add axis labels and a title
## complete the argument(s) to facet_wrap()
ggplot(monona, aes(x = ff_x)) +
  geom_histogram(center = 180, binwidth = 9,
                 color = "black", fill = "blue") +
  xlab("Days After June 30th") +
  ylab("Count") +
  ggtitle("Lake Monona Freeze Duartion, 1855-2022") +
  facet_wrap(vars(period50))
```



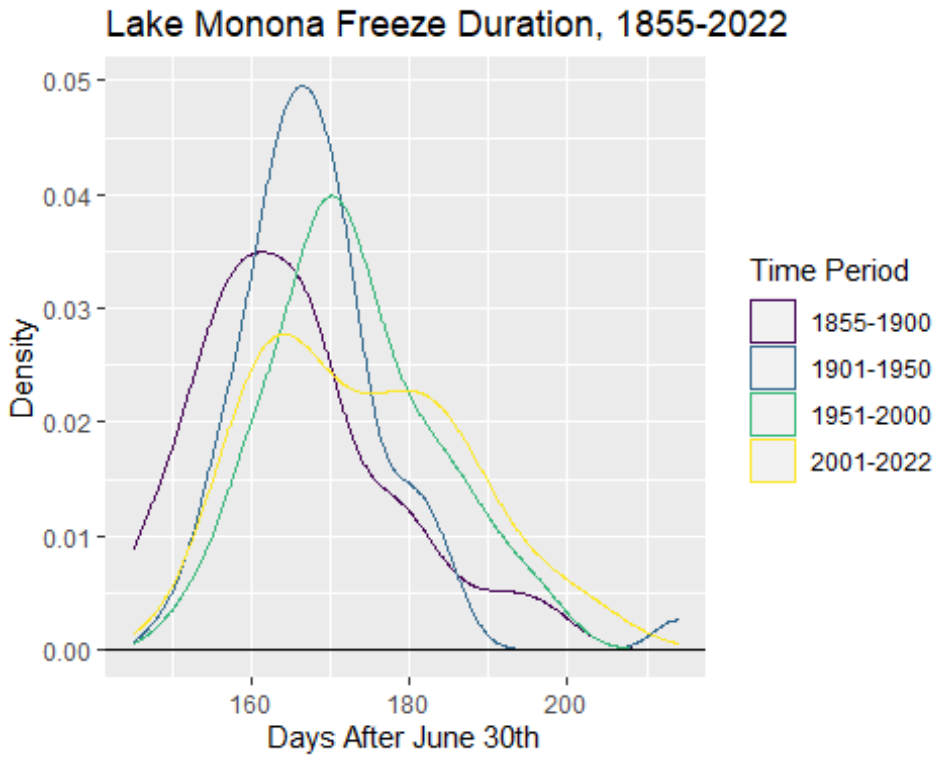

```
## histograms, stacked
##
## change the binwidth to an appropriate value
## add axis labels and a title
## complete the argument(s) to facet_grid()
##
ggplot(monona, aes(x = ff_x)) +
  geom_histogram(center = 180, binwidth = 6,
                 color = "black", fill = "blue") +
  xlab("Days After June 30th") +
  ylab("Count") +
  ggtitle("Lake Monona Freeze Duration, 1855-2022") +
  facet_grid(vars(period50))
```




```
## side by side boxplots
##
ggplot(monona, aes(x = as.character(period50), y=ff_x)) +
  geom_boxplot(color = "black", fill = "red",varwidth=TRUE) +
  xlab("Time Periods (By Years)")+
  ylab("Days After June 30th") +
  ggtitle("Lake Monona Freeze Duration, 1855-2022")
```




```
## Overlapping density plots
## You want a different color for each group of period50
## Add an appropriate aesthetic mapping
#
ggplot(monona, aes(x = ff_x, color=period50)) +
  geom_density() +
  geom_hline(yintercept = 0) +
  scale_color_viridis_d()+
  xlab("Days After June 30th") +
  ylab("Density") +
  ggtitle("Lake Monona Freeze Duration, 1855-2022") +
  labs(colour= "Time Period")
```



10

From the graphs in the previous problem, provide pros and cons for each. Which one or ones make it easiest to compare features about how the distribution of dates that Lake Monona is closed by ice varies among these time periods?

- **Wrapped Histograms**:
  - Pros: Helps us mainly with seeing a frequency distribution because we can see how many times Lake Monona's first freeze in days after June 30th for each time period in a simplistic format. The benefit over the stack histograms is ascetically it is easier to see if there is a skew in the data since it is less stretched out. This shows the positive and negative relationship better with the skew's.
  - Cons: Histograms are limited to showcasing group values and not the exact values for the data. Also, since they are based off one variable they are limited in the sense of not being able to showcase a central tendency. Lastly, compared to stacked histograms it is less effective in seeing difference in count for each time period.
- **Stacked Histograms**:
  - Pros: The basis of the histograms have the same pro as wrapped histograms because they showcase the Lake Monona's first freeze in days after June 30th for each time period in the same format. However, the wrapped histograms help present the data for the days after June 30th first freeze more efficient

because each bar from different time period represents the same days after June 30th for first freeze. This shows the count better.
  - Cons: The cons of the foundation for histograms as wrapped histograms. Also, the stacked histograms do not do a good job of showing the neg or positive relationship (with skews) because they are more stretched out.
- **Side-by-side Box Plots**:
  - Pros: Box plots show us a median, upper quartile, lower quartile, and outliers for the data. Also, the side by side feature helps us use those measurements to judge change in the median, LQ, and UQ pairs with them being side by side to create estimates of change over the time periods.
  - Cons: The main cons of box plot is they showcase the tails of the distribution which are most of the time are the least concrete data values and distribution reflection is limited because they mainly focus on central tendency and values that reflect off it.
- **Overlayed Density Plots**:
  - Pros:The benefit of density plots is they show the distribution shape the most efficiently because there is no constraint on the distribution. Also, it helps see the peak (highest density of data) more effective for each time period.
  - Cons: Downfall of density plots is there are not as simplistic to read like a histogram or box plot because you have to know higher density is a higher concentration of data points and vice versa.