Midterm 2 W24

Elisabeth Sellinger

2024-02-27

Instructions

Answer the following questions and complete the exercises in RMarkdown. Please embed all of your code and push your final work to your repository. Your code must be organized, clean, and run free from errors. Remember, you must remove the `#` for any included code chunks to run. Be sure to add your name to the author header above.

Your code must knit in order to be considered. If you are stuck and cannot answer a question, then comment out your code and knit the document. You may use your notes, labs, and homework to help you complete this exam. Do not use any other resources- including AI assistance.

Don’t forget to answer any questions that are asked in the prompt. Some questions will require a plot, but others do not- make sure to read each question carefully.

For the questions that require a plot, make sure to have clearly labeled axes and a title. Keep your plots clean and professional-looking, but you are free to add color and other aesthetics.

Be sure to follow the directions and upload your exam on Gradescope.

Background

In the `data` folder, you will find data about shark incidents in California between 1950-2022. The data (https://catalog.data.gov/dataset/shark-incident-database-california-56167) are from: State of California- Shark Incident Database.

Load the libraries

```
library("tidyverse")
library("janitor")
library("naniar")
```

Load the data

Run the following code chunk to import the data.

```
sharks <- read_csv("data/SharkIncidents_1950_2022_220302.csv") %>% clean_names()
```

Questions

1. (1 point) Start by doing some data exploration using your preferred function(s). What is the structure of the data? Where are the missing values and how are they represented?

```
glimpse(sharks)
```

```
## Rows: 211
## Columns: 16
## $ incident_num     <chr> "1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "1…
## $ month            <dbl> 10, 5, 12, 2, 8, 4, 10, 5, 6, 7, 10, 11, 4, 5, 5, 8, …
## $ day              <dbl> 8, 27, 7, 6, 14, 28, 12, 7, 14, 28, 4, 10, 24, 19, 21…
## $ year             <dbl> 1950, 1952, 1952, 1955, 1956, 1957, 1958, 1959, 1959,…
## $ time             <chr> "12:00", "14:00", "14:00", "12:00", "16:30", "13:30",…
## $ county           <chr> "San Diego", "San Diego", "Monterey", "Monterey", "Sa…
## $ location         <chr> "Imperial Beach", "Imperial Beach", "Lovers Point", "…
## $ mode             <chr> "Swimming", "Swimming", "Swimming", "Freediving", "Sw…
## $ injury           <chr> "major", "minor", "fatal", "minor", "major", "fatal",…
## $ depth            <chr> "surface", "surface", "surface", "surface", "surface"…
## $ species          <chr> "White", "White", "White", "White", "White", "White",…
## $ comment          <chr> "Body Surfing, bit multiple times on leg, thigh and b…
## $ longitude        <chr> "-117.1466667", "-117.2466667", "-122.05", "-122.15",…
## $ latitude         <dbl> 32.58833, 32.58833, 36.62667, 36.62667, 35.13833, 35.…
## $ confirmed_source <chr> "Miller/Collier, Coronado Paper, Oceanside Paper", "G…
## $ wfl_case_number  <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, N…
```

```
sharks %>%
map_df(~ sum(is.na(.)))
```

```
## # A tibble: 1 × 16
##   incident_num month   day  year  time county location  mode injury depth
##          <int> <int> <int> <int> <int>  <int>    <int> <int>  <int> <int>
## 1            0     0     0     0     7      0        0     0      0     0
## # ℹ 6 more variables: species <int>, comment <int>, longitude <int>,
## #   latitude <int>, confirmed_source <int>, wfl_case_number <int>
```

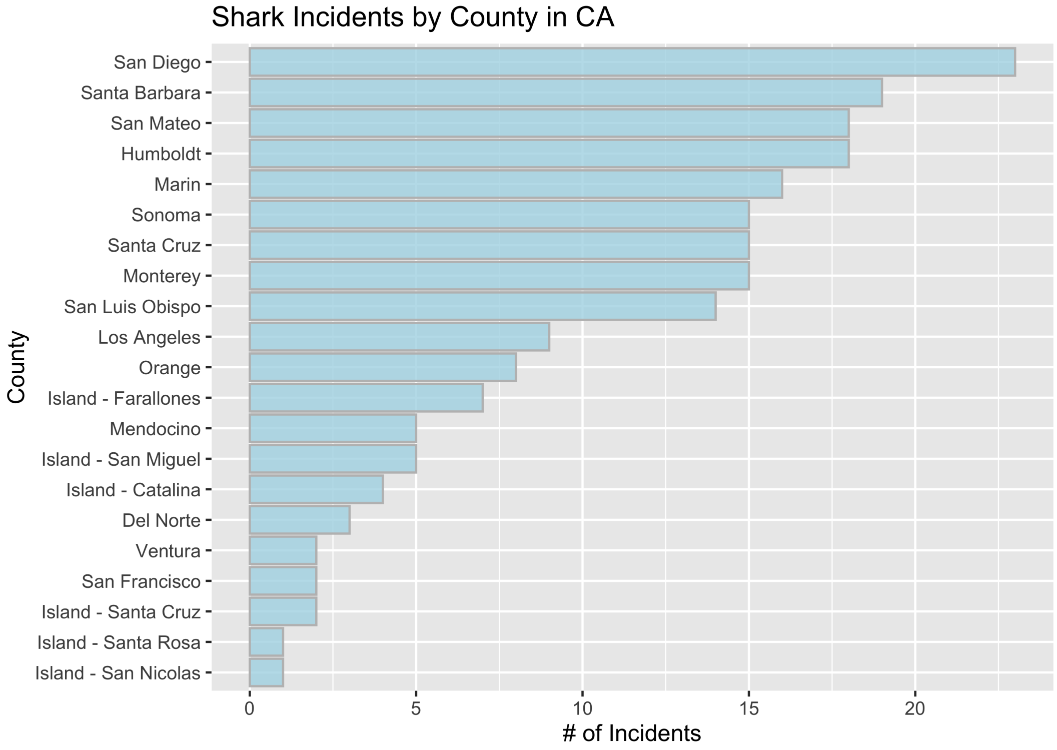
The structure of the data contains characters and factors with 211 rows and 16 columns. The NA's are represented with 'NA' and are the most prevelant in the 'wfl_case_number' column.

2. (1 point) Notice that there are some incidents identified as "NOT COUNTED". These should be removed from the data because they were either not sharks, unverified, or were provoked. It's OK to replace the `sharks` object.

```
sharks <- sharks %>%
  filter(incident_num != "NOT COUNTED")
```

3. (3 points) Are there any "hotspots" for shark incidents in California? Make a plot that shows the total number of incidents per county. Which county has the highest number of incidents?

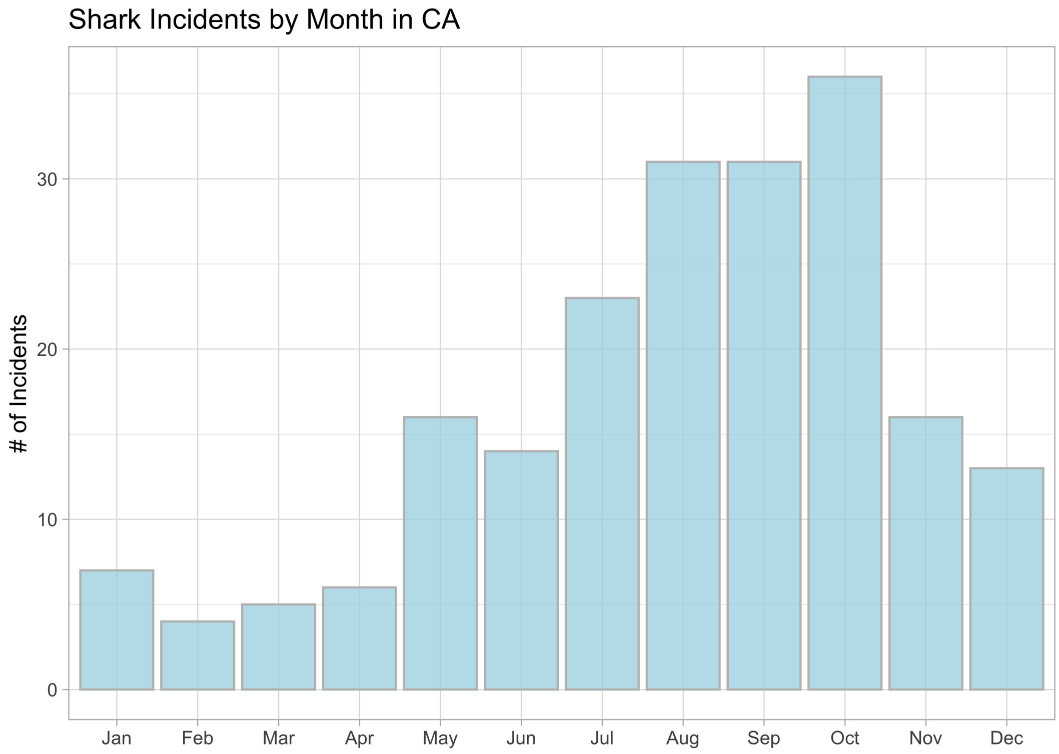
```
sharks %>%
  count(county) %>%
  ggplot(aes(x = reorder(county, n), y = n)) +
  geom_col(alpha = .8, fill = "lightblue", color = "grey") +
  coord_flip() +
  labs(title = "Shark Incidents by County in CA",
       x = "County",
       y = "# of Incidents")
```



San Diego has the highest number of incidents making it a “hotspot”.

4. (3 points) Are there months of the year when incidents are more likely to occur? Make a plot that shows the total number of incidents by month. Which month has the highest number of incidents?

```
sharks %>%
  count(month) %>%
  ggplot(aes(x=month, y=n))+
  geom_col(alpha = .8, fill = "lightblue", color = "grey")+
  labs(title="Shark Incidents by Month in CA",
       x=NULL,
       y="# of Incidents")+
  scale_x_discrete(limits=c("Jan","Feb","Mar", "Apr", "May", "Jun", "Jul", "Aug", "Sep",
"Oct", "Nov", "Dec"))+
  theme_light()
```



October has the highest number of incidents.

5. (3 points) How do the number and types of injuries compare by county? Make a table (not a plot) that shows the number of injury types by county. Which county has the highest number of fatalities?

```
sharks %>%
  group_by(county, injury) %>%
  summarise(num_incidents = n())
```

```
## `summarise()` has grouped output by 'county'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 57 × 3
## # Groups:   county [21]
##    county               injury num_incidents
##    <chr>                <chr>          <int>
##  1 Del Norte            minor              2
##  2 Del Norte            none               1
##  3 Humboldt             major              7
##  4 Humboldt             minor              2
##  5 Humboldt             none               9
##  6 Island - Catalina    minor              1
##  7 Island - Catalina    none               3
##  8 Island - Farallones major              7
##  9 Island - San Miguel fatal              1
## 10 Island - San Miguel major              2
## # ℹ 47 more rows
```

```
sharks %>%
  group_by(county, injury) %>%
  summarise(num_incidents = n()) %>%
  filter(injury == "fatal") %>%
  arrange(desc(num_incidents))
```

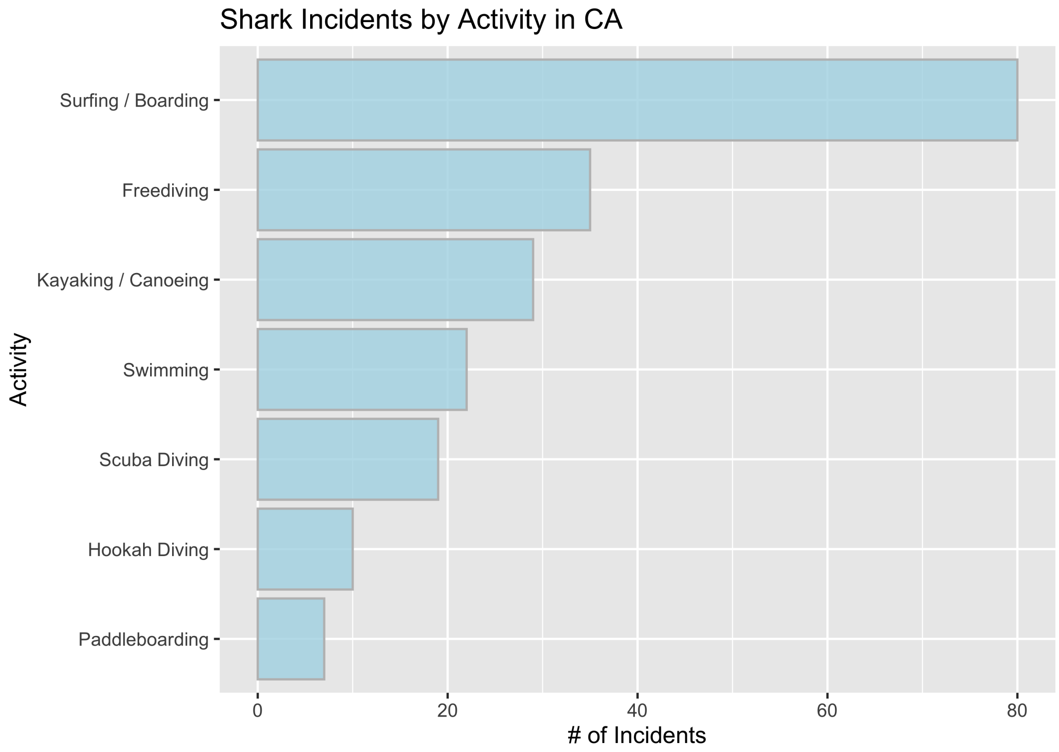
```
## `summarise()` has grouped output by 'county'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 10 × 3
## # Groups:   county [10]
##    county              injury num_incidents
##    <chr>               <chr>          <int>
##  1 San Luis Obispo     fatal              3
##  2 Monterey            fatal              2
##  3 San Diego           fatal              2
##  4 Santa Barbara       fatal              2
##  5 Island - San Miguel fatal              1
##  6 Los Angeles         fatal              1
##  7 Mendocino           fatal              1
##  8 San Francisco       fatal              1
##  9 San Mateo           fatal              1
## 10 Santa Cruz          fatal              1
```

San Luis Obispo has the highest number of fatalities at 3.

6. (2 points) In the data, `mode` refers to a type of activity. Which activity is associated with the highest number of incidents?

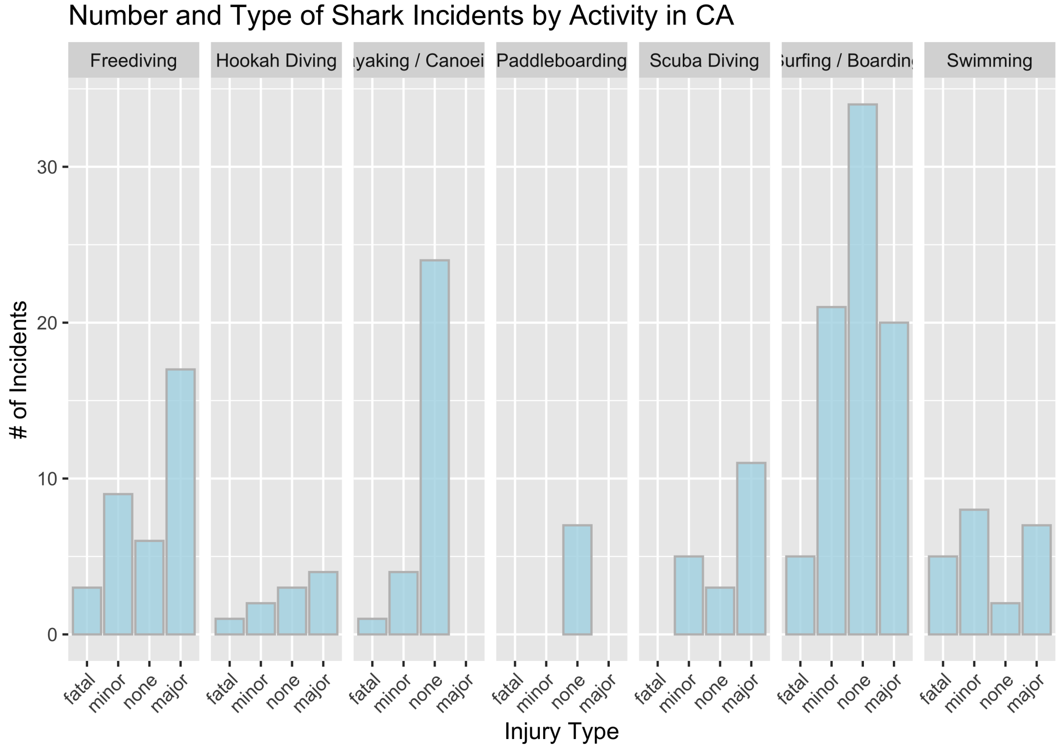
```
sharks %>%
  count(mode) %>%
  ggplot(aes(x = reorder(mode, n), y = n)) +
  geom_col(alpha = .8, fill = "lightblue", color = "grey") +
  coord_flip() +
  labs(title = "Shark Incidents by Activity in CA",
       x = "Activity",
       y = "# of Incidents")
```



Surfing/boarding has the highest number of incidents

7. (4 points) Use faceting to make a plot that compares the number and types of injuries by activity. (hint: the x axes should be the type of injury)

```
sharks %>%
  group_by(mode, injury) %>%
  summarise(num_incidents = n(), .groups = "keep") %>%
  ggplot(aes(x = reorder(injury, num_incidents), y = num_incidents)) +
  geom_col(alpha = .8, fill = "lightblue", color = "grey") +
  facet_grid(.~mode) +
  theme(axis.text.x = element_text(angle = 45, hjust=1)) +
  labs(title = "Number and Type of Shark Incidents by Activity in CA",
       x = "Injury Type",
       y = "# of Incidents")
```



8. (1 point) Which shark species is involved in the highest number of incidents?

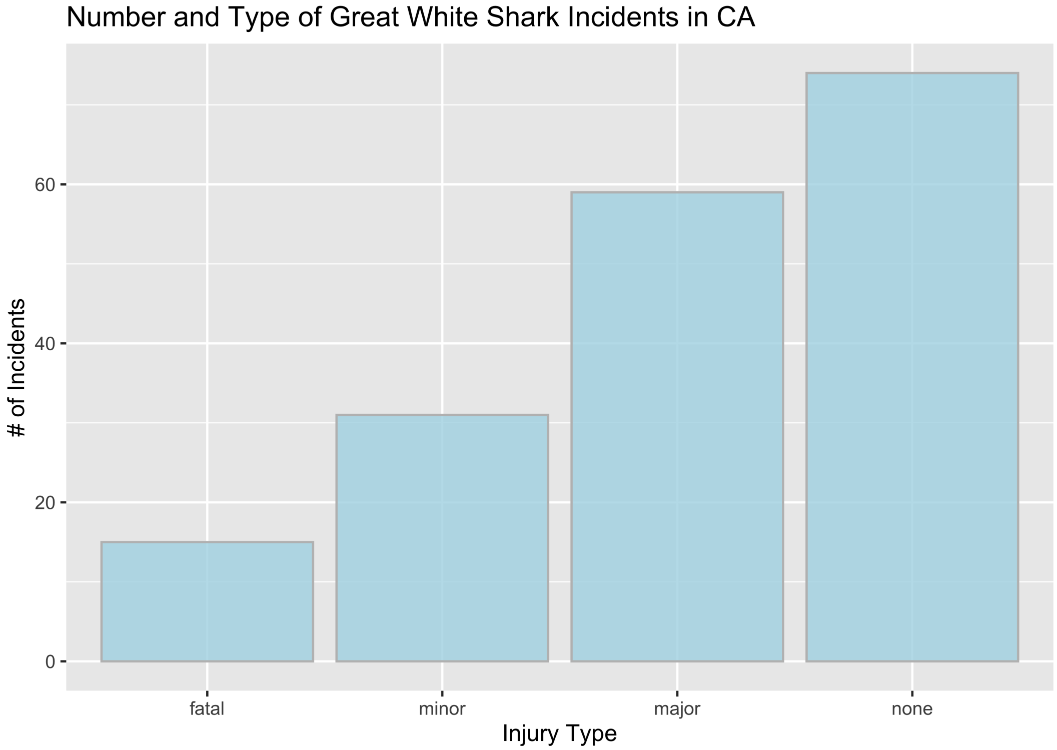
```
sharks %>%
  group_by(species) %>%
  summarise(num_incidents = n()) %>%
  arrange(desc(num_incidents))
```

```
## # A tibble: 8 × 2
##   species    num_incidents
##   <chr>              <int>
## 1 White                179
## 2 Unknown               13
## 3 Hammerhead             3
## 4 Blue                   2
## 5 Leopard                2
## 6 Salmon                 1
## 7 Sevengill              1
## 8 Thresher               1
```

Great White sharks have the most incidents

9. (3 points) Are all incidents involving Great White's fatal? Make a plot that shows the number and types of injuries for Great White's only.

```
sharks %>%
  filter(species == "White") %>%
  group_by(injury) %>%
  summarise(num_incidents = n()) %>%
  ggplot(aes(x = reorder(injury, num_incidents), y = num_incidents)) +
  geom_col(alpha = .8, fill = "lightblue", color = "grey") +
  labs(title = "Number and Type of Great White Shark Incidents in CA",
          x = "Injury Type",
          y = "# of Incidents")
```



No, not all Great White incidents are fatal, in fact, most of them aren’t.

Background

Let’s learn a little bit more about Great White sharks by looking at a small dataset that tracked 20 Great White’s in the Fallaron Islands. The data (https://link.springer.com/article/10.1007/s00227-007-0739-4) are from: Weng et al. (2007) Migration and habitat of white sharks (*Carcharodon carcharias*) in the eastern Pacific Ocean.

Load the data

```
white_sharks <- read_csv("data/White sharks tracked from Southeast Farallon Island, CA,
USA, 1999 2004.csv", na = c("?", "n/a")) %>% clean_names()
```

10. (1 point) Start by doing some data exploration using your preferred function(s). What is the structure of the data? Where are the missing values and how are they represented?

```
glimpse(white_sharks)
```

```
## Rows: 20
## Columns: 10
## $ shark           <chr> "1-M", "2-M", "3-M", "4-M", "5-F", "6-M", "7-F", "8-M"…
## $ tagging_date    <chr> "19-Oct-99", "30-Oct-99", "16-Oct-00", "5-Nov-01", "5-…
## $ total_length_cm <dbl> 402, 366, 457, 457, 488, 427, 442, 380, 450, 530, 427,…
## $ sex             <chr> "M", "M", "M", "M", "F", "M", "F", "M", "M", "F", NA, …
## $ maturity        <chr> "Mature", "Adolescent", "Mature", "Mature", "Mature", …
## $ pop_up_date     <chr> "2-Nov-99", "25-Nov-99", "16-Apr-01", "6-May-02", "19-…
## $ track_days      <dbl> 14, 26, 182, 182, 256, 275, 35, 60, 209, 91, 182, 240,…
## $ longitude       <dbl> -124.49, -125.97, -156.80, -141.47, -133.25, -138.83, …
## $ latitude        <dbl> 38.95, 38.69, 20.67, 26.39, 21.13, 26.50, 37.07, 34.93…
## $ comment         <chr> "Nearshore", "Nearshore", "To Hawaii", "To Hawaii", "O…
```

```
white_sharks %>%
map_df(~ sum(is.na(.)))
```

```
## # A tibble: 1 × 10
##   shark tagging_date total_length_cm   sex maturity pop_up_date track_days
##   <int>        <int>           <int> <int>    <int>       <int>      <int>
## 1     0            0               0     3        1           0          0
## # ℹ 3 more variables: longitude <int>, latitude <int>, comment <int>
```

The structure of the data contains characters and integers with 20 rows and 10 columns. The NA's are represented with 'NA' and there aren't very many.

11. (3 points) How do male and female sharks compare in terms of total length? Are males or females larger on average? Do a quick search online to verify your findings. (hint: this is a table, not a plot).

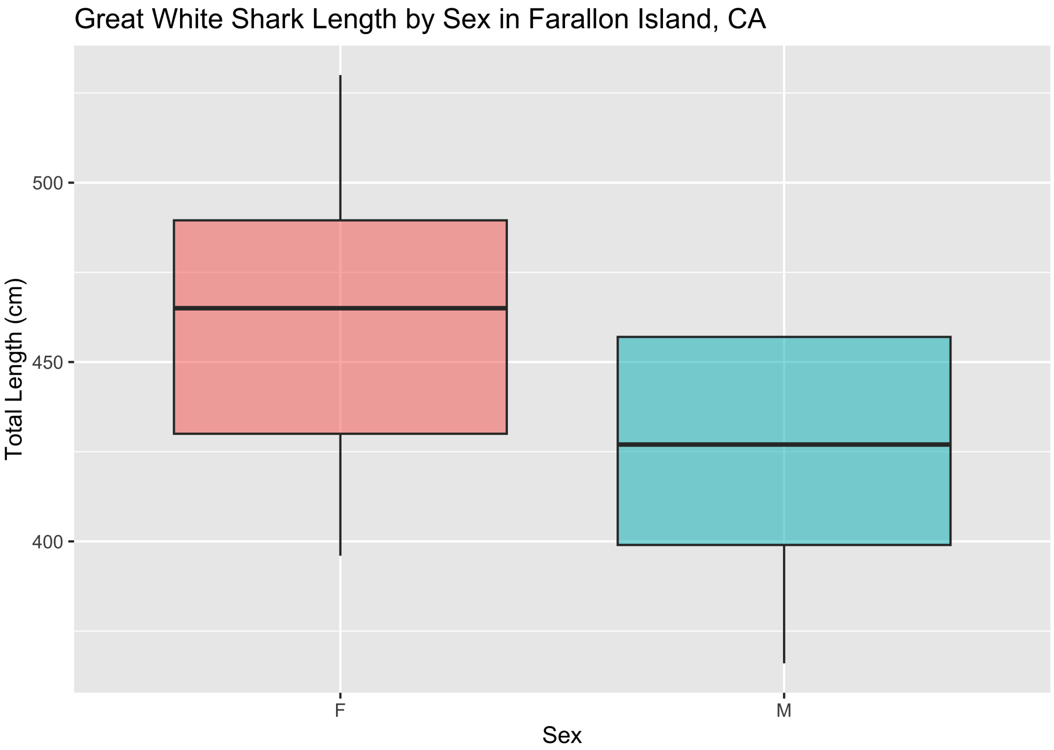
```
white_sharks %>%
  filter(sex != "NA") %>%
  group_by(sex) %>%
  summarise(mean_length = mean(total_length_cm))
```

```
## # A tibble: 2 × 2
##   sex   mean_length
##   <chr>       <dbl>
## 1 F            462
## 2 M            425.
```

On average, females are longer than males. This is still true after searching online to verify my findings.

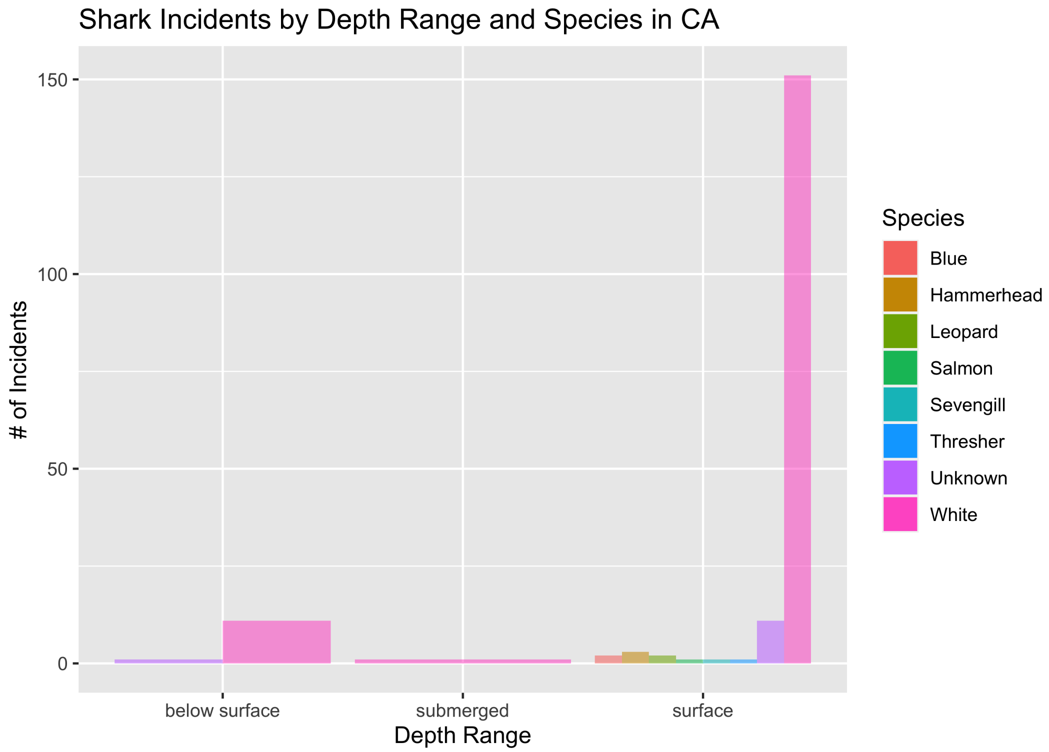
12. (3 points) Make a plot that compares the range of total length by sex.

```
white_sharks %>%
  filter(sex != "NA") %>%
  ggplot(aes(x = sex, y = total_length_cm, fill = sex, alpha = .5)) +
  geom_boxplot() +
  guides(fill = "none",
         alpha = "none") +
  labs(title = "Great White Shark Length by Sex in Farallon Island, CA",
       x = "Sex",
       y = "Total Length (cm)")
```



13. (2 points) Using the `sharks` or the `white_sharks` data, what is one question that you are interested in exploring? Write the question and answer it using a plot or table.
#### Where are most of the incidents occuring in the water column? Compare this by species.

```
sharks %>%
  mutate(depth_range = case_when(depth == "surface" ~ "surface",
                                   depth == "submerged" ~ "submerged",
                                   depth > 5 & depth <= 25 ~ "shallow",
                                   depth > 26 ~ "below surface")) %>%
  group_by(depth_range, species) %>%
  count(num_incidents = n()) %>%
  filter(depth_range != "N/A") %>%
  ggplot(aes(x = depth_range, y = n, fill = species, alpha = .8)) +
  geom_col(position = "dodge") +
  guides(alpha = "none") +
  labs(title = "Shark Incidents by Depth Range and Species in CA",
        x = "Depth Range",
       y = "# of Incidents",
       fill = "Species")
```



Most Incidents are at the surface and are by Great White Sharks