Midterm 2 W24

Jacques Mak

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 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…
```

```
anyNA(sharks)
```

```
## [1] TRUE
```

```
miss_var_summary(sharks)
```

```
## # A tibble: 16 × 3
##    variable         n_miss pct_miss
##    <chr>             <int>    <dbl>
##  1 wfl_case_number     202   95.7
##  2 time                  7    3.32
##  3 latitude              6    2.84
##  4 longitude             5    2.37
##  5 confirmed_source      1    0.474
##  6 incident_num          0    0
##  7 month                 0    0
##  8 day                   0    0
##  9 year                  0    0
## 10 county                0    0
## 11 location              0    0
## 12 mode                  0    0
## 13 injury                0    0
## 14 depth                 0    0
## 15 species               0    0
## 16 comment               0    0
```

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

```
## # A tibble: 202 × 16
##    incident_num month   day  year time    county     location mode  injury depth
##    <chr>        <dbl> <dbl> <dbl> <chr>   <chr>      <chr>    <chr> <chr>  <chr>
##  1 1               10     8  1950 12:00   San Diego  Imperia… Swim… major  surf…
##  2 2                5    27  1952 14:00   San Diego  Imperia… Swim… minor  surf…
##  3 3               12     7  1952 14:00   Monterey   Lovers … Swim… fatal  surf…
##  4 4                2     6  1955 12:00   Monterey   Pacific… Free… minor  surf…
##  5 5                8    14  1956 16:30   San Luis … Pismo B… Swim… major  surf…
##  6 6                4    28  1957 13:30   San Luis … Morro B… Swim… fatal  surf…
##  7 7               10    12  1958 Unknown San Diego  Coronad… Swim… major  surf…
##  8 8                5     7  1959 17:30   San Franc… Baker B… Swim… fatal  surf…
##  9 9                6    14  1959 17:00   San Diego  La Jolla Free… fatal  surf…
## 10 10               7    28  1959 19:30   San Diego  La Jolla Free… minor  surf…
## # ℹ 192 more rows
## # ℹ 6 more variables: species <chr>, comment <chr>, longitude <chr>,
## #   latitude <dbl>, confirmed_source <chr>, wfl_case_number <chr>
```

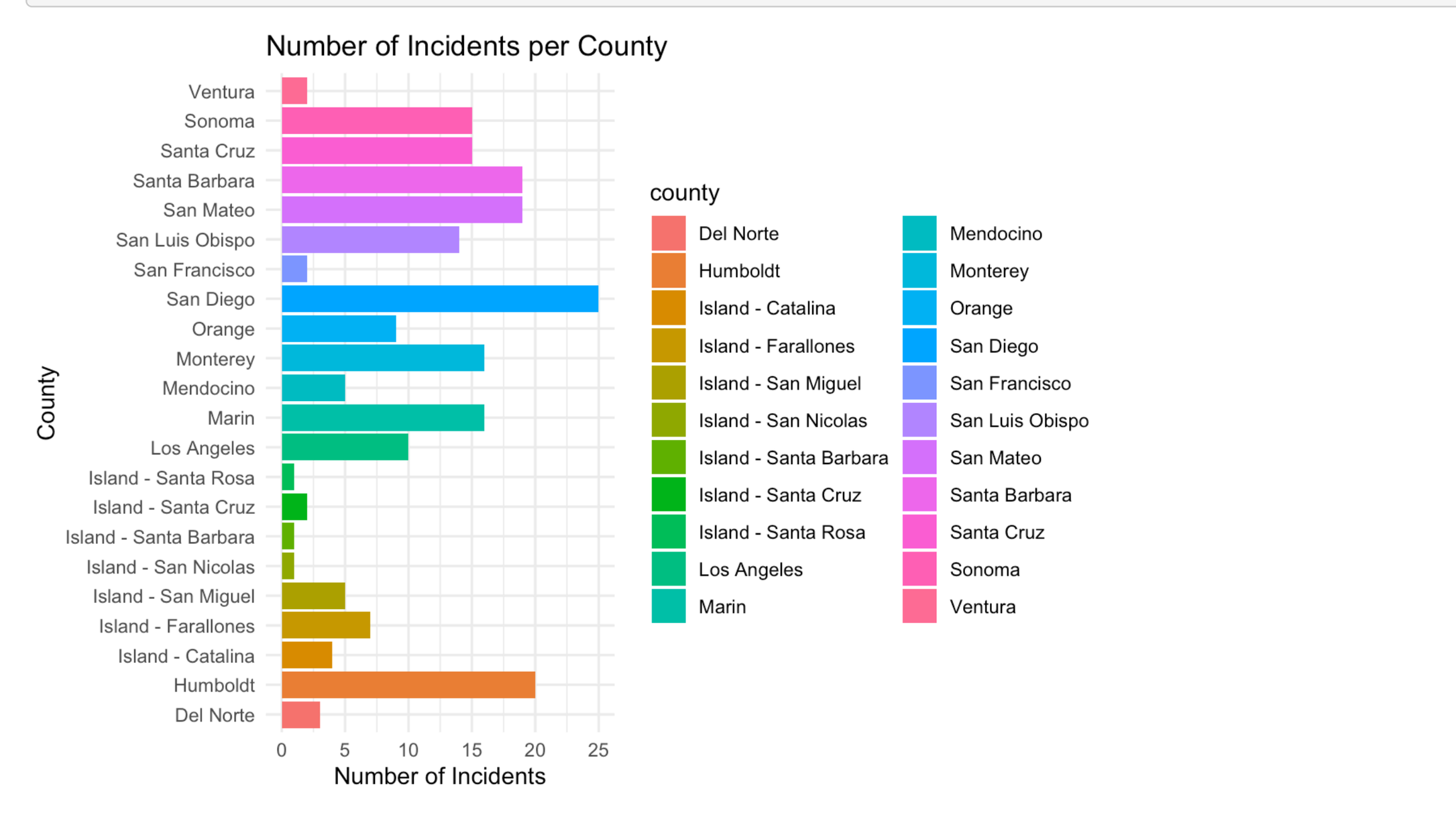
3. (3 points) Are there any "hotspots" for shark incidents in California? Make a plot that shows the total number of incidents per county. Which county has the highest number of incidents?

```
sharks %>%
  filter(incident_num!="NOT COUNTED") %>%
 count(county)%>%
  arrange(desc(n))
```

```
## # A tibble: 21 × 2
##    county              n
##    <chr>           <int>
##  1 San Diego          23
##  2 Santa Barbara      19
##  3 Humboldt           18
##  4 San Mateo          18
##  5 Marin              16
##  6 Monterey           15
##  7 Santa Cruz         15
##  8 Sonoma             15
##  9 San Luis Obispo    14
## 10 Los Angeles         9
## # ℹ 11 more rows
```

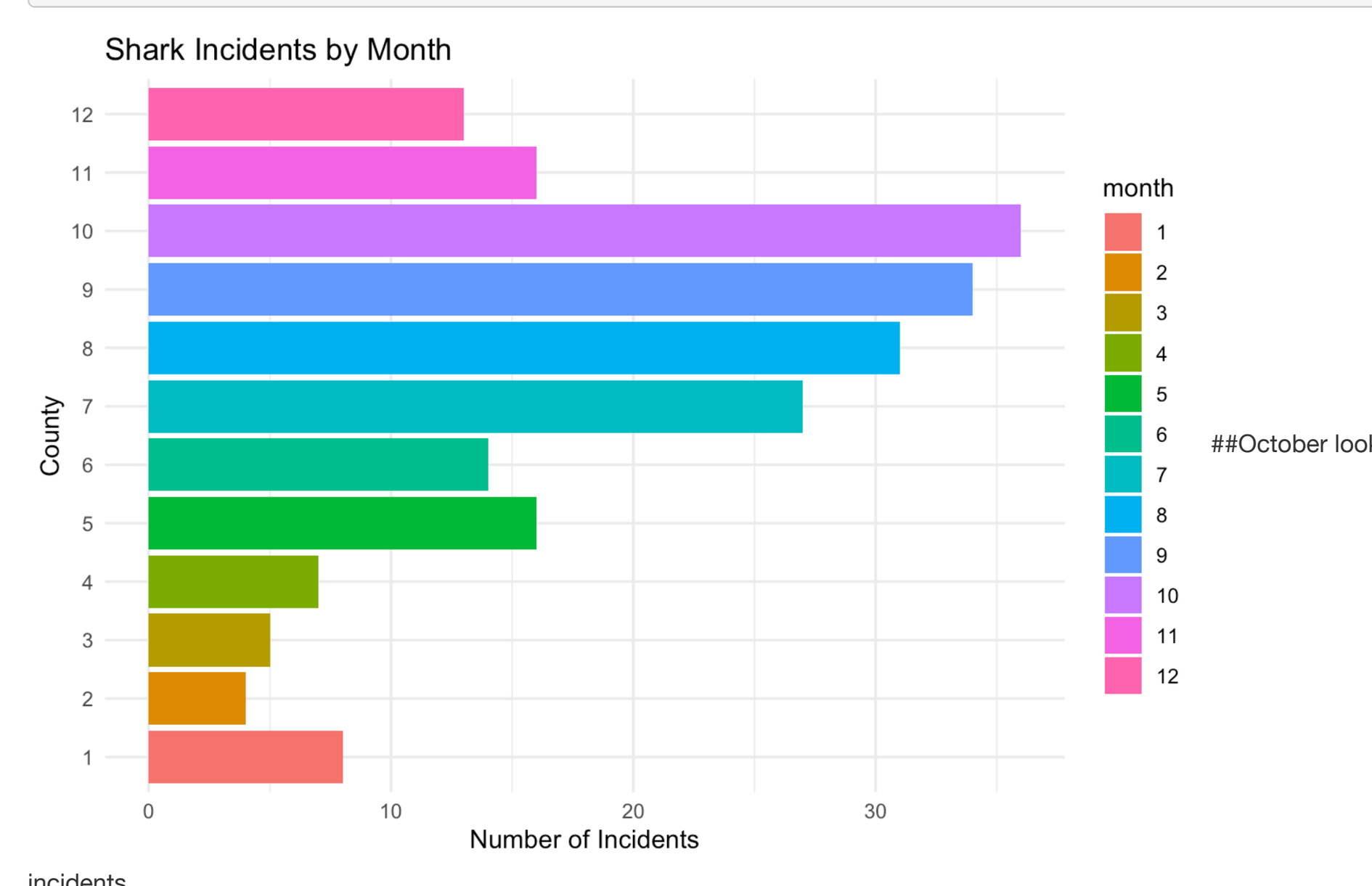
##San Diego has the largest total number of incident.

```
sharks %>%
  ggplot(aes(fill = county, x = county)) +
  geom_bar() +
  coord_flip() +
  labs(title = "Number of Incidents per County",
       y = "Number of Incidents",
       x = "County") +theme_minimal()
```



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$month <- as.factor(sharks$month)
sharks %>%
  ggplot(aes(fill = month, x = month)) +
  geom_bar() +
  coord_flip() +
  labs(title = "Shark Incidents by Month",
       y = "Number of Incidents",
       x = "County") +theme_minimal()
```



##October looks like a hotspot of the 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(total =  n()) %>%
  pivot_wider(names_from = injury,
              values_from = total) %>%
  arrange(desc(fatal))
```

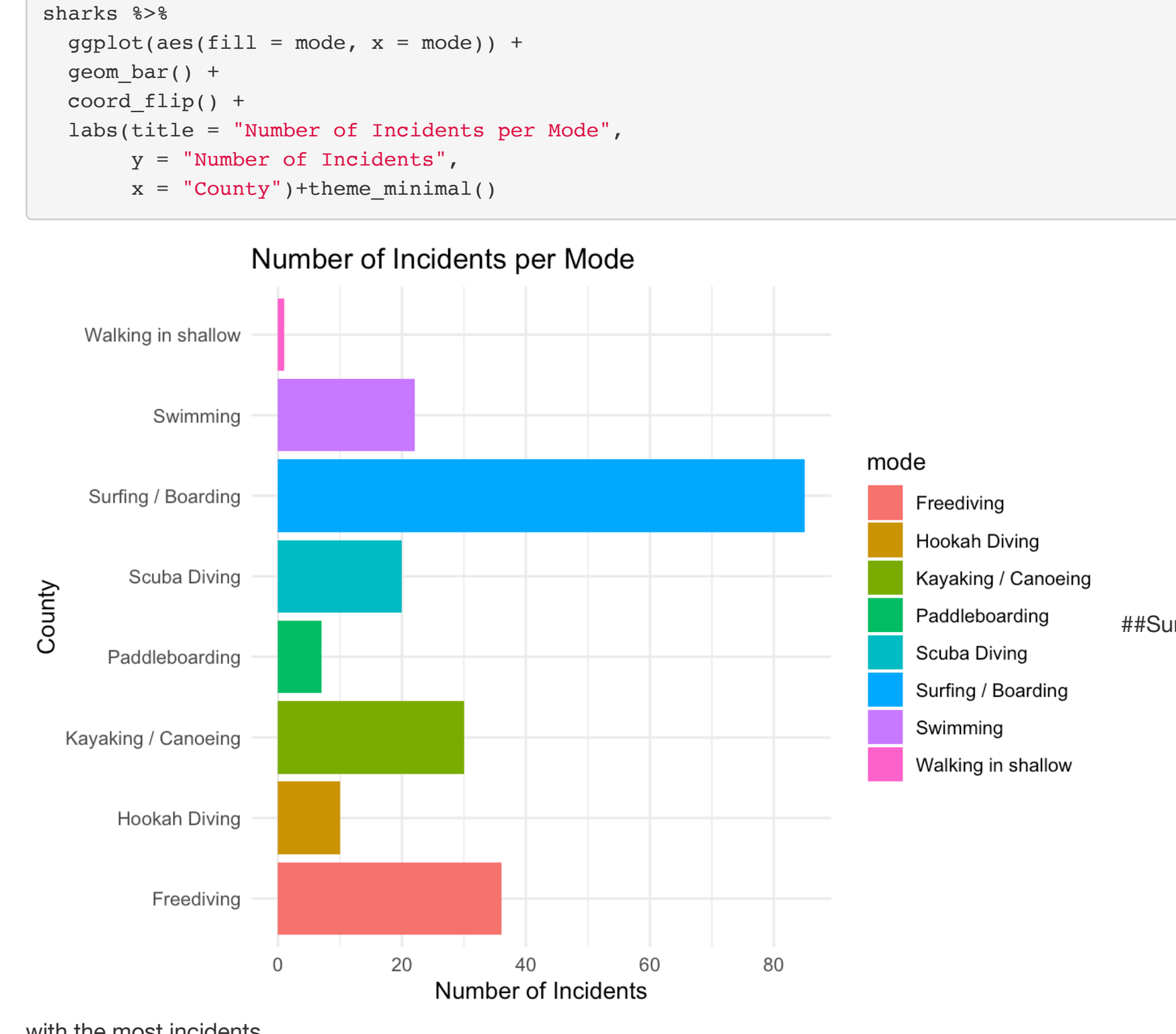
```
## `summarise()` has grouped output by 'county'. You can override using the
## `.groups` argument.
```

```
## # A tibble: 22 × 8
## # Groups:   county [22]
##    county              minor  none major `minor*` `none*` fatal `major*`
##    <chr>               <int> <int> <int>    <int>   <int> <int>    <int>
##  1 San Luis Obispo         1     7     3       NA      NA     3       NA
##  2 Monterey                2     3     8       NA      NA     2        1
##  3 San Diego               8     9     4        1       1     2       NA
##  4 Santa Barbara           6     9     2       NA      NA     2       NA
##  5 Island - San Miguel     2    NA     2       NA      NA     1       NA
##  6 Los Angeles             6     2    NA       NA      NA     1        1
##  7 Mendocino               1    NA     3       NA      NA     1       NA
##  8 San Francisco          NA     1    NA       NA      NA     1       NA
##  9 San Mateo               4    12     1        1      NA     1       NA
## 10 Santa Cruz              3     8     3       NA      NA     1       NA
## # ℹ 12 more rows
```

##San luis Obispo has to most fatal incidents.

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

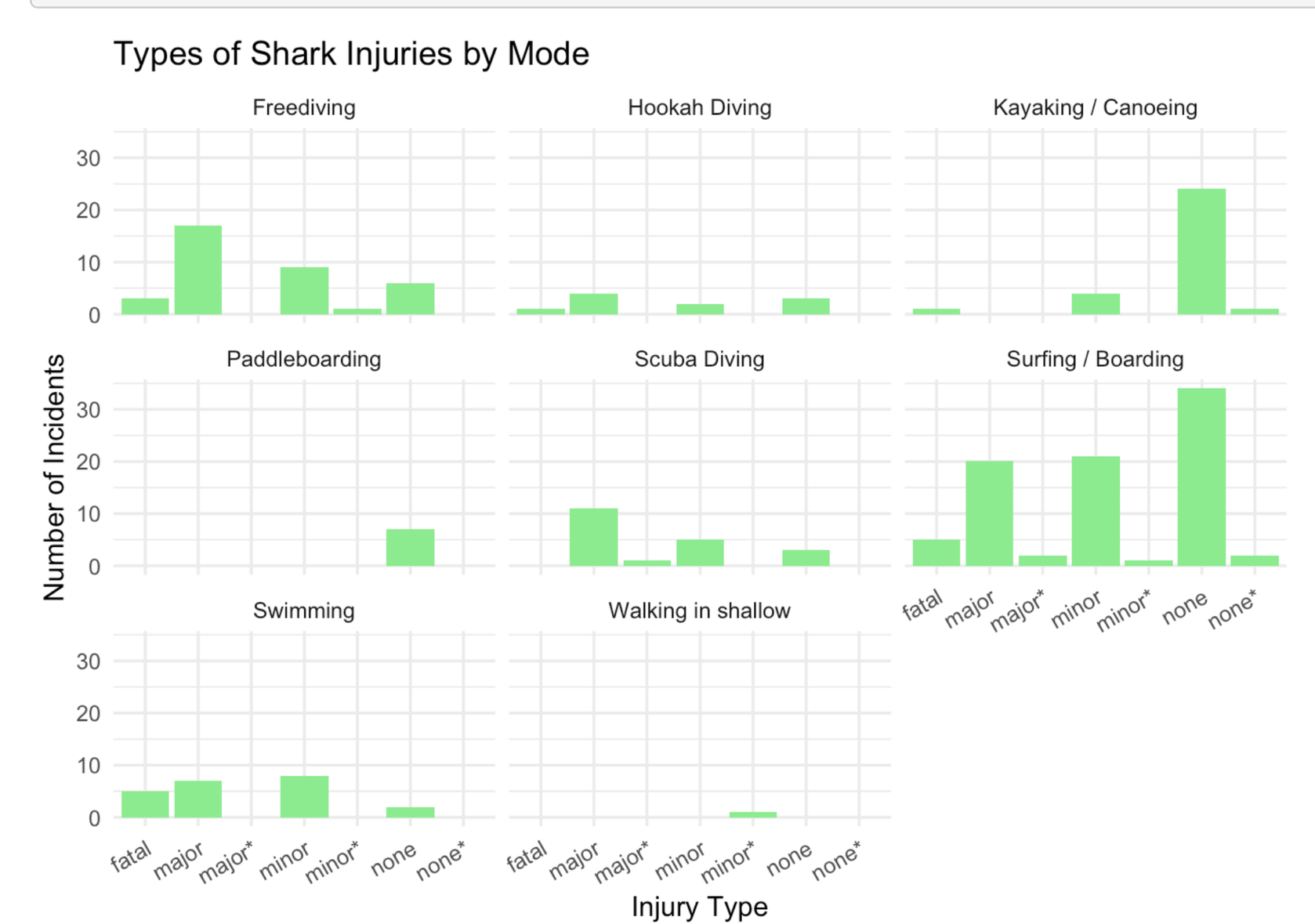
```
sharks %>%
  ggplot(aes(fill = mode, x = mode)) +
  geom_bar() +
  coord_flip() +
  labs(title = "Number of Incidents per Mode",
       y = "Number of Incidents",
       x = "County")+theme_minimal()
```



##Surfing/ Boardinf is associating with the most incidents.

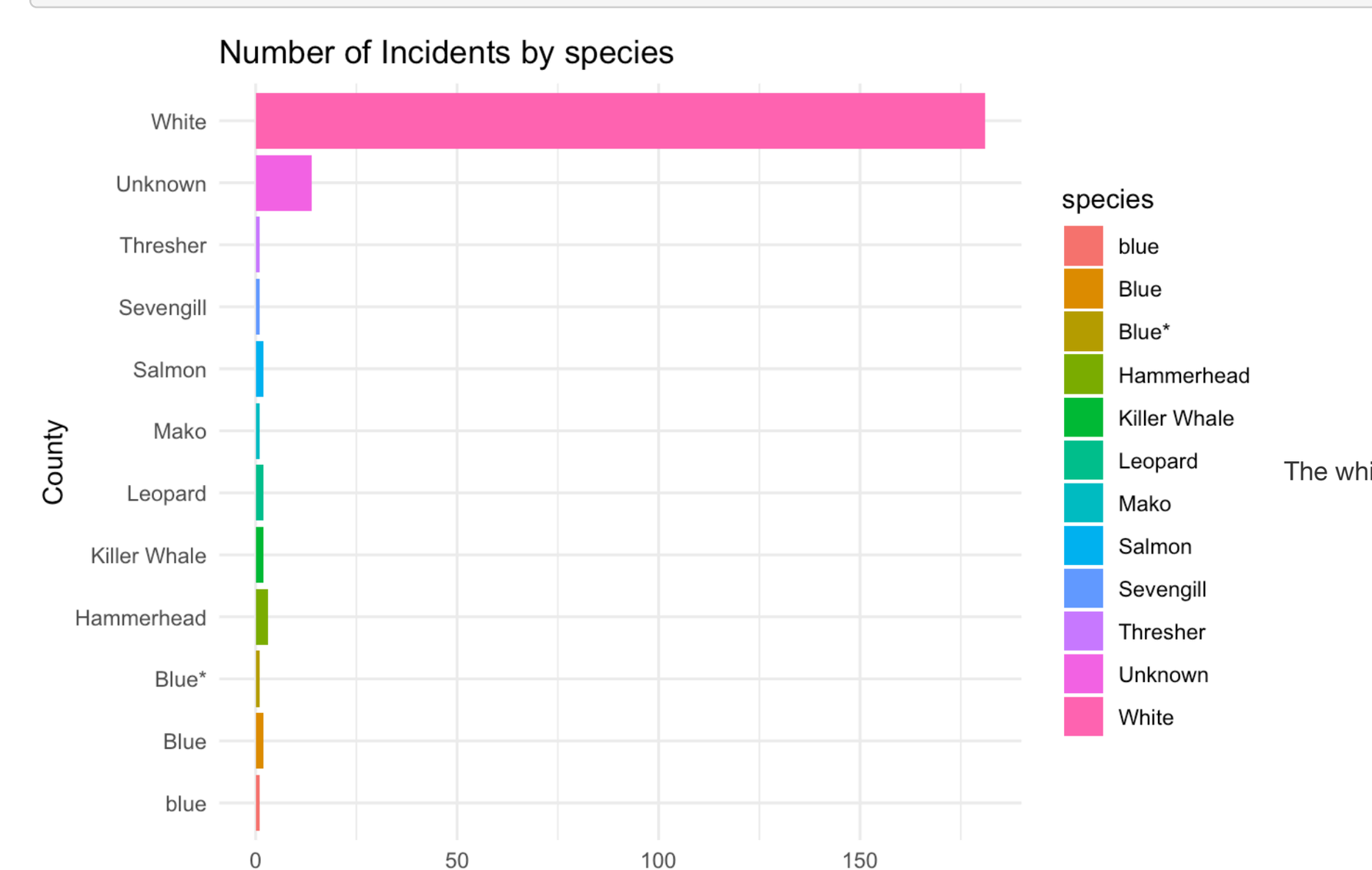
7. (4 points) Use faceting to make a plot that compares the number and types of injuries by activity. (hint: the x axes should be the type of injury)

```
sharks %>%
  group_by(mode, injury) %>%
  ggplot(aes(x = injury)) +
  geom_bar(fill = "light green") +
  facet_wrap(~mode) +
  labs(title = "Types of Shark Injuries by Mode",
       x = "Injury Type",
       y = "Number of Incidents") +
  theme_minimal()+ theme(axis.text.x = element_text(angle = 30, hjust=1))
```



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

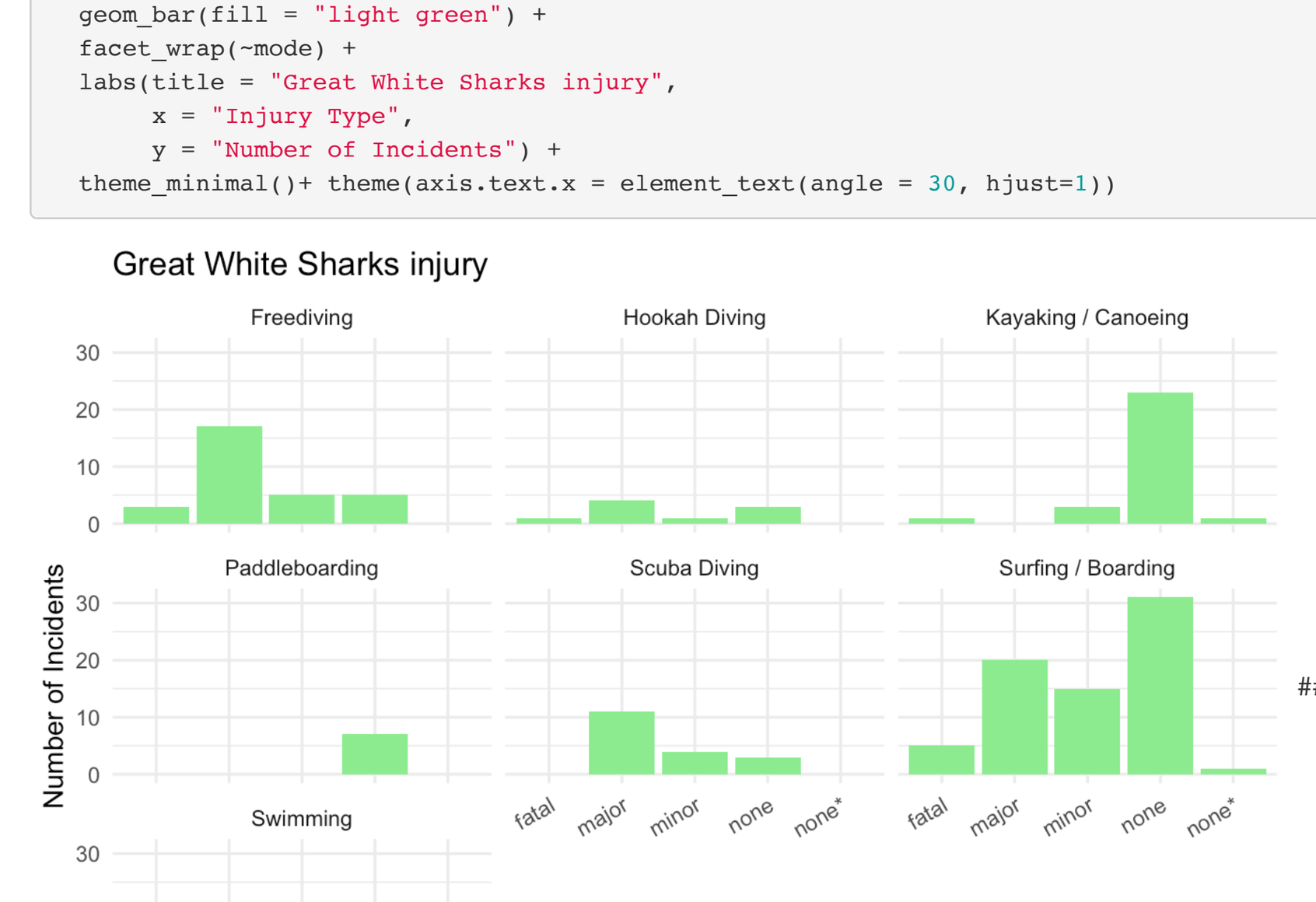
```
sharks %>%
  ggplot(aes(fill = species, x = species)) +
  geom_bar() +
  coord_flip() +
  labs(title = "Number of Incidents by species",
       y = "Number of Incidents",
       x = "County") +theme_minimal()
```



The white shark has contributes 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(mode, injury) %>%
  ggplot(aes(x = injury)) +
  geom_bar(fill = "light green") +
  facet_wrap(~mode) +
  labs(title = "Great White Sharks injury",
       x = "Injury Type",
       y = "Number of Incidents") +
  theme_minimal()+ theme(axis.text.x = element_text(angle = 30, hjust=1))
```



##Not all of the Great White Sharks incidents are fatal.

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 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?

```
summary(white_sharks)
```

```
##     shark           tagging_date       total_length_cm     sex
##  Length:20          Length:20          Min.   :360.0   Length:20
##  Class :character   Class :character   1st Qu.:400.5   Class :character
##  Mode  :character   Mode  :character   Median :434.5   Mode  :character
##                                        Mean   :436.1
##                                        3rd Qu.:457.0
##                                        Max.   :530.0
##
##    maturity         pop_up_date          track_days      longitude
##  Length:20          Length:20          Min.   : 14.0   Min.   :-156.8
##  Class :character   Class :character   1st Qu.: 85.0   1st Qu.:-137.8
##  Mode  :character   Mode  :character   Median :182.0   Median :-133.2
##                                        Mean   :166.8   Mean   :-120.3
##                                        3rd Qu.:216.8   3rd Qu.:-124.3
##                                        Max.   :367.0   Max.   : 131.7
##                                                        NA's   :1
##     latitude       comment
##  Min.   :20.67   Length:20
##  1st Qu.:22.48   Class :character
##  Median :26.39   Mode  :character
##  Mean   :28.24
##  3rd Qu.:36.00
##  Max.   :38.95
##  NA's   :1
```

```
miss_var_summary(white_sharks)
```

```
## # A tibble: 10 × 3
##    variable        n_miss pct_miss
##    <chr>            <int>    <dbl>
##  1 sex                  3       15
##  2 maturity             1        5
##  3 longitude            1        5
##  4 latitude             1        5
##  5 shark                0        0
##  6 tagging_date         0        0
##  7 total_length_cm      0        0
##  8 pop_up_date          0        0
##  9 track_days           0        0
## 10 comment              0        0
```

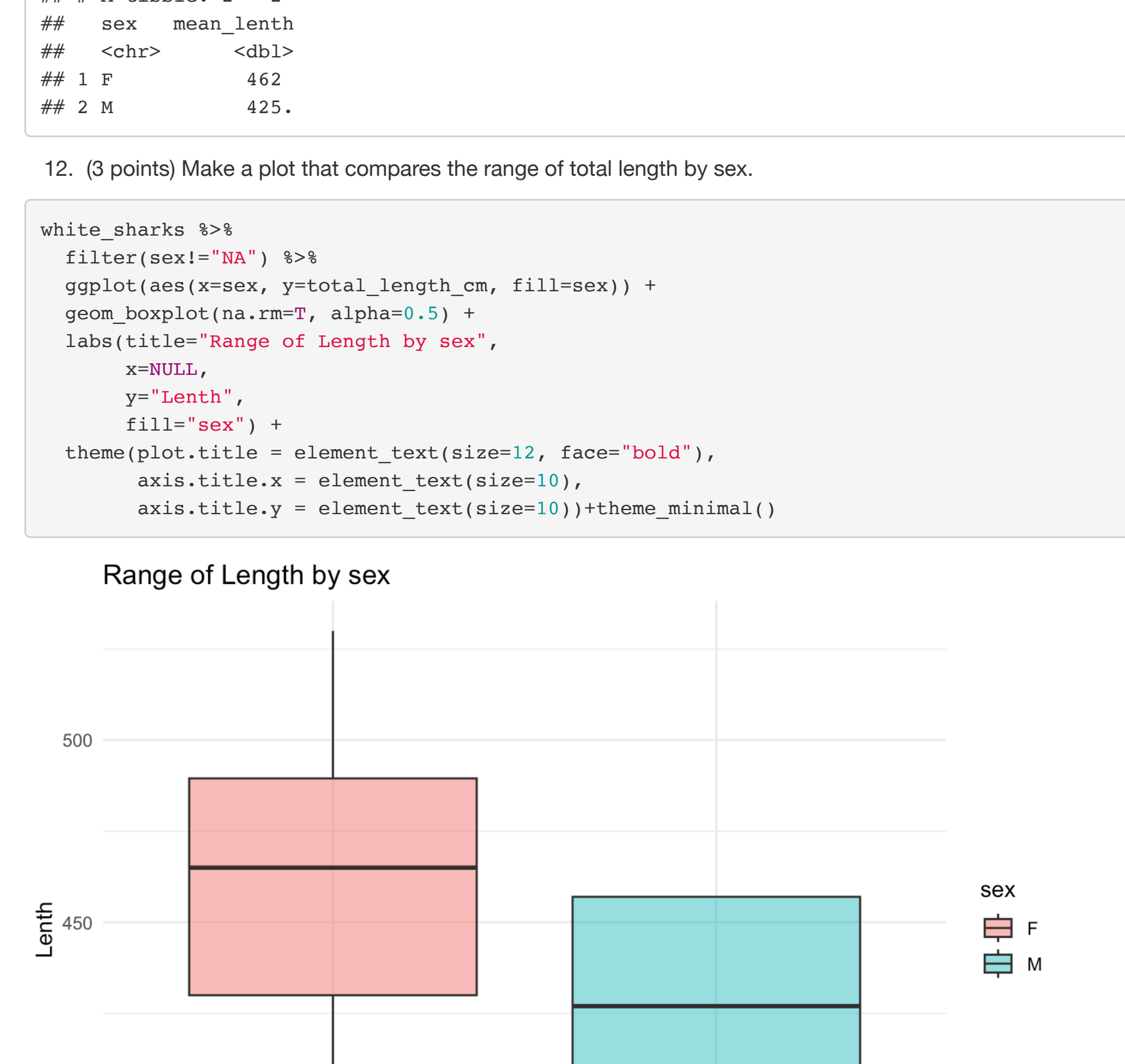
11. (3 points) How do male and female sharks compare in terms of total length? Are males or females larger on average? Do a quick search online to verify your findings. (hint: this is a table, not a plot).

```
white_sharks %>%
  filter(sex!="NA") %>%
  group_by(sex) %>%
  summarize(mean_lenth=mean(total_length_cm, na.rm=T))
```

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

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)) +
  geom_boxplot(na.rm=T, alpha=0.5) +
  labs(title="Range of Length by sex",
       x=NULL,
       y="Lenth",
       fill="sex") +
  theme(plot.title = element_text(size=12, face="bold"),
        axis.title.x = element_text(size=10),
        axis.title.y = element_text(size=10))+theme_minimal()
```



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.

What are the number and types of injuries specifically in Sonoma?

```
sharks %>%
 filter(county=="Sonoma")%>%
  group_by(mode, injury) %>%
  ggplot(aes(x = injury)) +
  geom_bar(fill = "light green") +
  facet_wrap(~mode) +
  labs(title = "Sonoma injury",
       x = "Injury Type",
       y = "Number of Incidents") +
  theme_minimal()+ theme(axis.text.x = element_text(angle = 30, hjust=1))
```