Midterm 2 W24

Spencer Bryan

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")
library(ggplot2)
```

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?

```
head(sharks)
```

```
## # A tibble: 6 × 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 Obi… Pismo B… Swim… major  surf…
## 6 6                4    28  1957 13:30 San Luis Obi… Morro B… Swim… fatal  surf…
## # ℹ 6 more variables: species <chr>, comment <chr>, longitude <chr>,
## #   latitude <dbl>, confirmed_source <chr>, wfl_case_number <chr>
```

```
summary(sharks)
```

```
##  incident_num          month             day             year
##  Length:211         Min.   : 1.000   Min.   : 1.00   Min.   :1950
##  Class :character   1st Qu.: 6.000   1st Qu.: 7.50   1st Qu.:1985
##  Mode  :character   Median : 8.000   Median :18.00   Median :2004
##                     Mean   : 7.858   Mean   :16.54   Mean   :1998
##                     3rd Qu.:10.000   3rd Qu.:25.00   3rd Qu.:2014
##                     Max.   :12.000   Max.   :31.00   Max.   :2022
##
##      time              county            location            mode
##  Length:211         Length:211         Length:211         Length:211
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##     injury             depth             species            comment
##  Length:211         Length:211         Length:211         Length:211
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   longitude           latitude     confirmed_source   wfl_case_number
##  Length:211         Min.   :32.59   Length:211         Length:211
##  Class :character   1st Qu.:34.04   Class :character   Class :character
##  Mode  :character   Median :36.70   Mode  :character   Mode  :character
##                     Mean   :36.36
##                     3rd Qu.:38.18
##                     Max.   :41.56
##                     NA's   :6
```

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.

   I am removing unverified shark attack observations from the data I will be analyzing.

```
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?

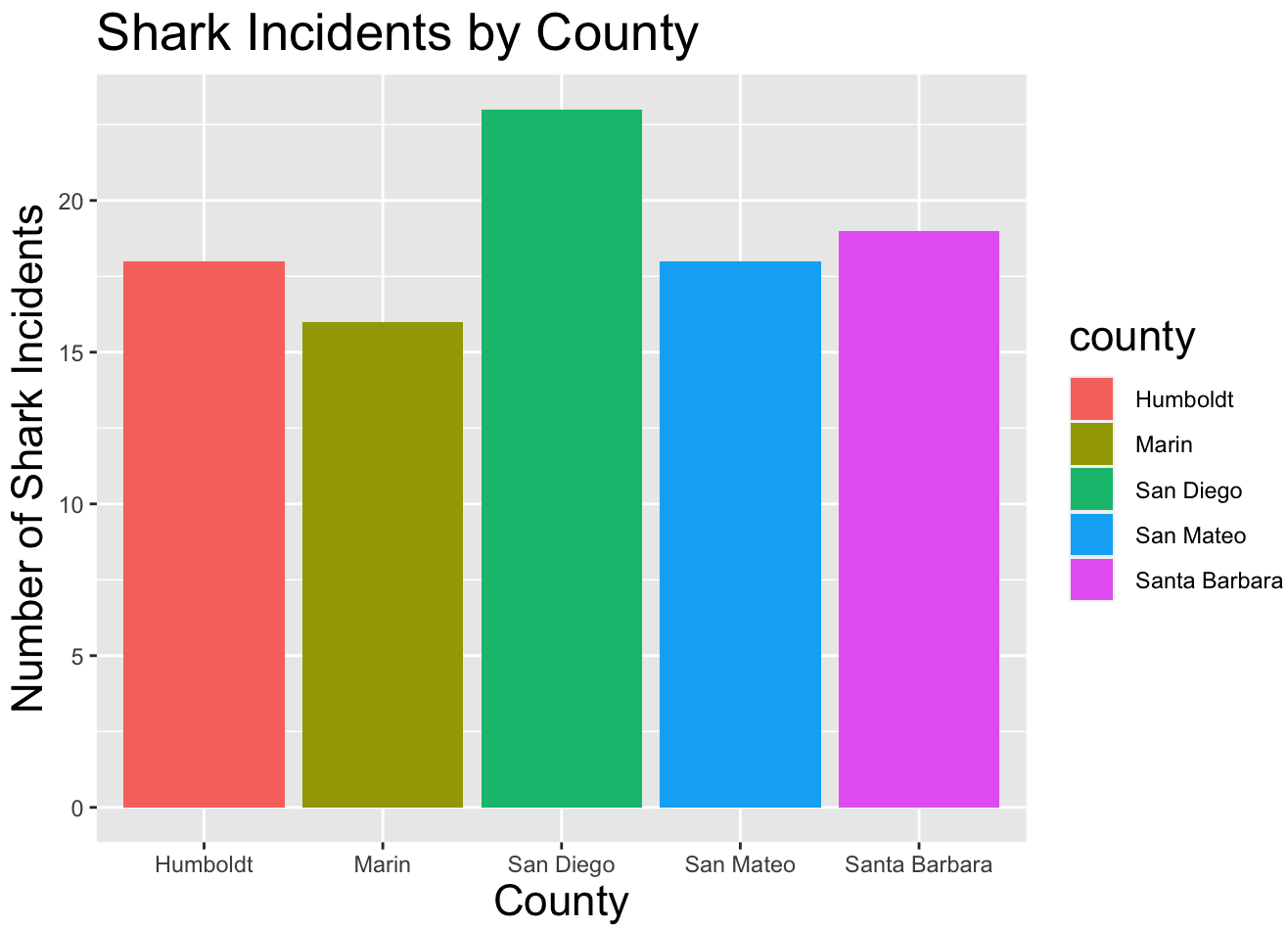
   Numerical data in the form of a table

```
sharks %>%
  group_by(county) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  select(county, number_incidents) %>%
  arrange(-number_incidents)
```

```
## # A tibble: 21 × 2
##    county          number_incidents
##    <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
```

```
The table shows that San Deigo, Santa Barbara, Humboldt, Marin, and Moneteray counties h
ave the highest number of shark incidents.
Graphical representaion of the table
```

```
sharks %>%
  group_by(county) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  top_n(5, number_incidents) %>%
  ggplot(aes(x= county, y= number_incidents, fill = county)) +
  geom_col() +
  labs(title = "Shark Incidents by County",
       x= "County",
       y= "Number of Shark Incidents") +
  theme(title = element_text(size= rel(1.5), hjust= 0.5))
```

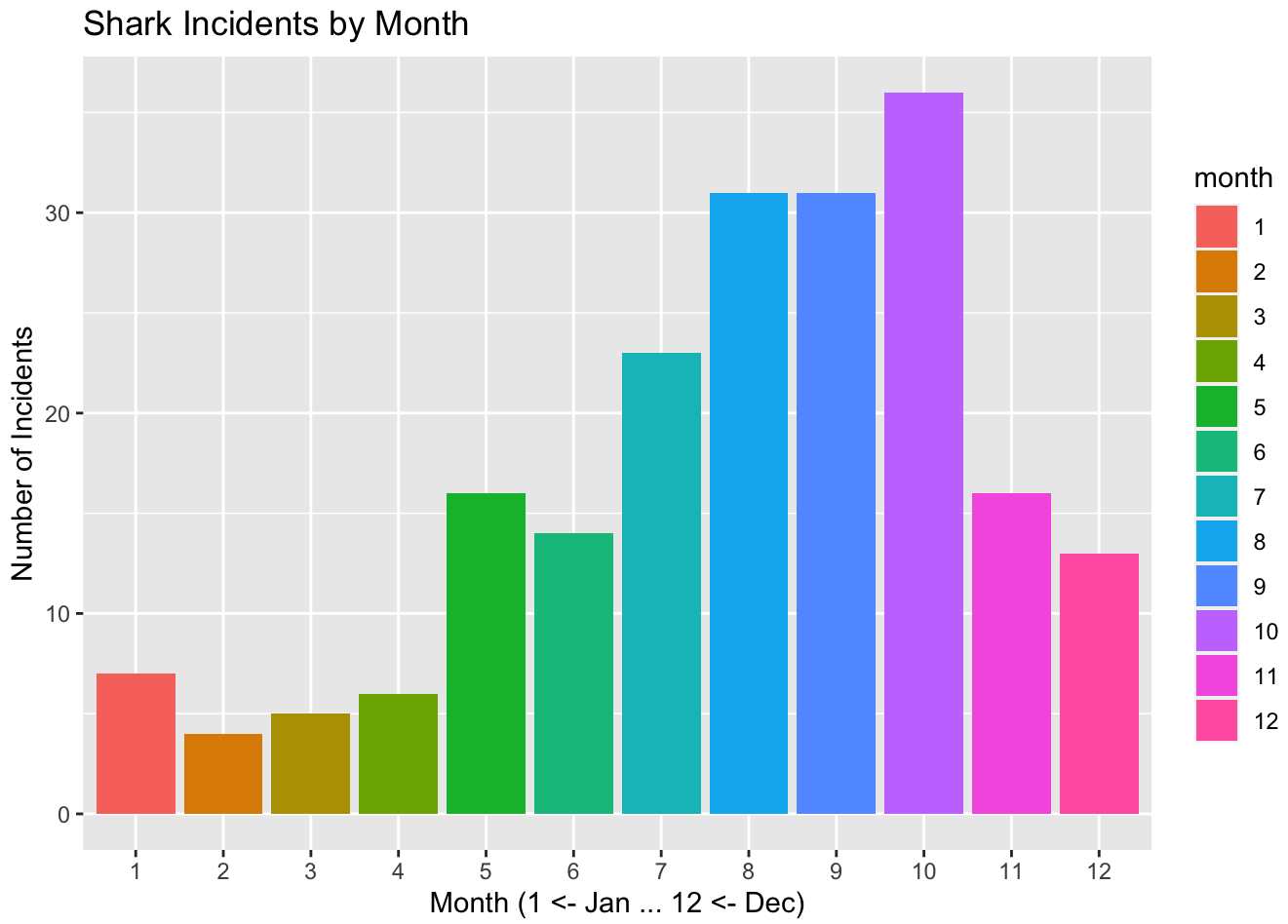


4. (3 points) Are there months of the year when incidents are more likely to occur? Make a plot that shows the total number of incidents by month. Which month has the highest number of incidents?

   Must have the month as a factor to be able to group using it.

```
sharks$month <- as.factor(sharks$month)
```

```
sharks %>%
  group_by(month) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  ggplot(aes(x= month, y= number_incidents, fill = month)) +
  geom_col() +
  labs(title = "Shark Incidents by Month",
      x= "Month (1 <- Jan ... 12 <- Dec)",
      y= "Number of Incidents")
```



August-October are the most likely months to have a shark related incident, with October having the hightest number of incidences.

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(number_injury = n()) %>%
  filter(injury == "fatal") %>%
  top_n(10, number_injury) %>%
  arrange(-number_injury)
```

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

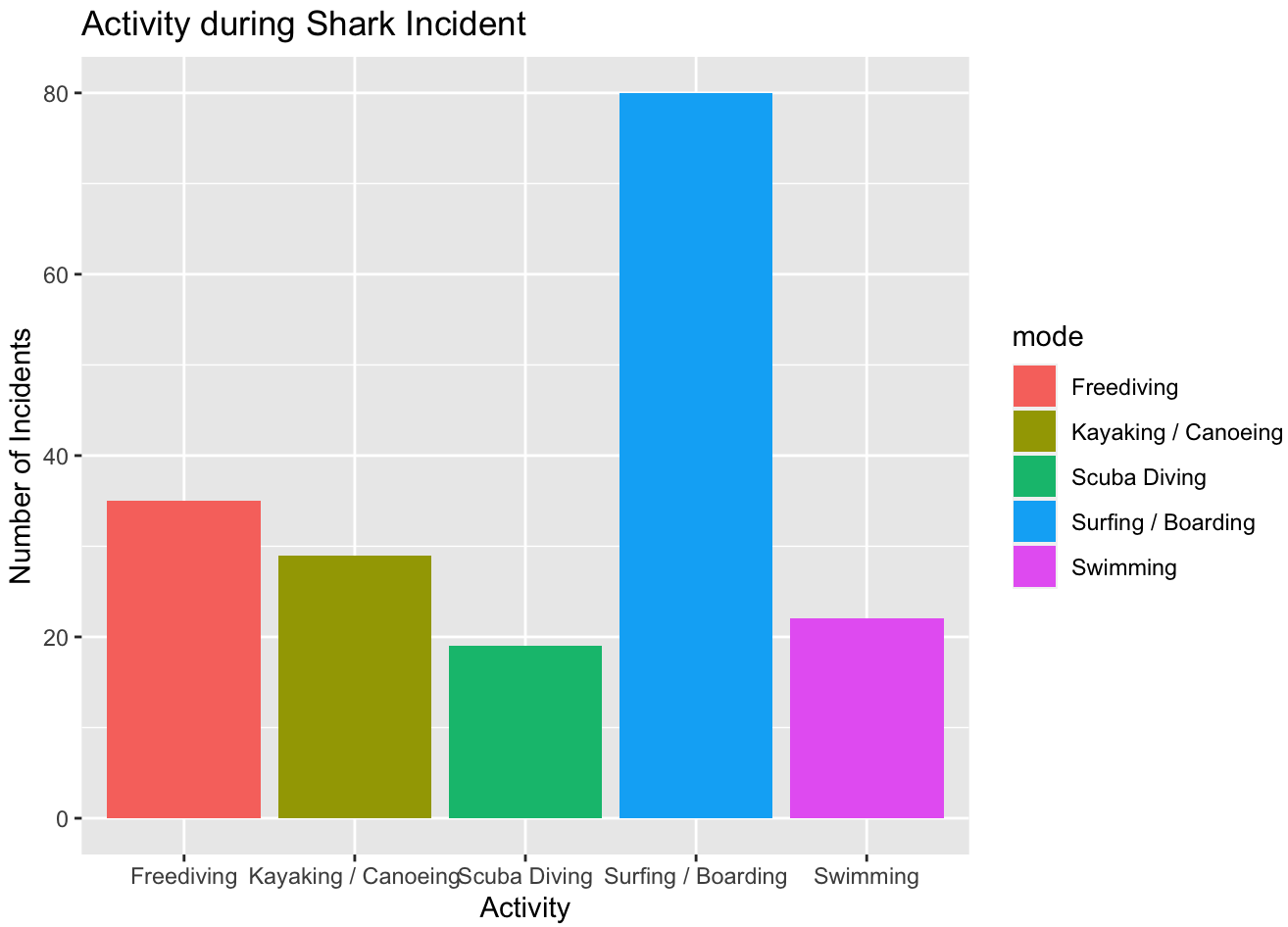
```
## # A tibble: 10 × 3
## # Groups:   county [10]
##    county              injury number_injury
##    <chr>               <chr>          <int>
##  1 San Luis Obispo     fatal              3
##  2 Monterey            fatal              2
##  3 San Diego           fatal              2
##  4 Santa Barbara       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
```

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

```
sharks %>%
  group_by(mode) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  arrange(-number_incidents)
```

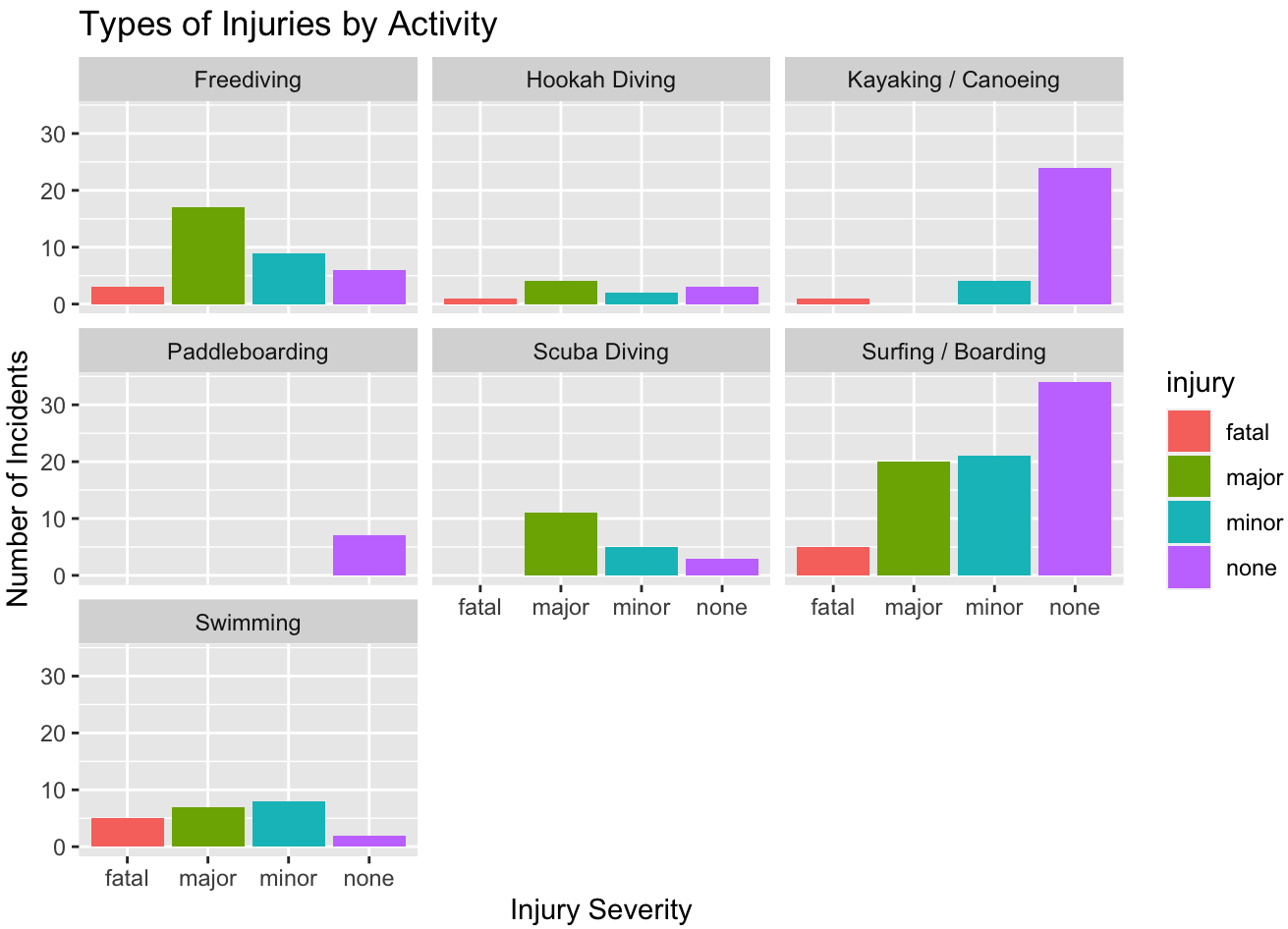
```
## # A tibble: 7 × 2
##   mode                number_incidents
##   <chr>                          <int>
## 1 Surfing / Boarding                80
## 2 Freediving                        35
## 3 Kayaking / Canoeing               29
## 4 Swimming                          22
## 5 Scuba Diving                      19
## 6 Hookah Diving                     10
## 7 Paddleboarding                     7
```

```
sharks %>%
  group_by(mode) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  top_n(5, number_incidents) %>%
  ggplot(aes(x= mode, y= number_incidents, fill = mode)) +
  geom_col() +
  labs(title = "Activity during Shark Incident",
       x= "Activity",
       y= "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 %>%
  ggplot(aes(injury, fill= injury)) +
  geom_bar() +
  facet_wrap(~mode) +
  labs(title = "Types of Injuries by Activity",
       x= "Injury Severity",
       y= "Number of Incidents")
```




```
Surfing has the highest number of incidents, but freediving has the highest proportion o
f major injuries per incident.
```

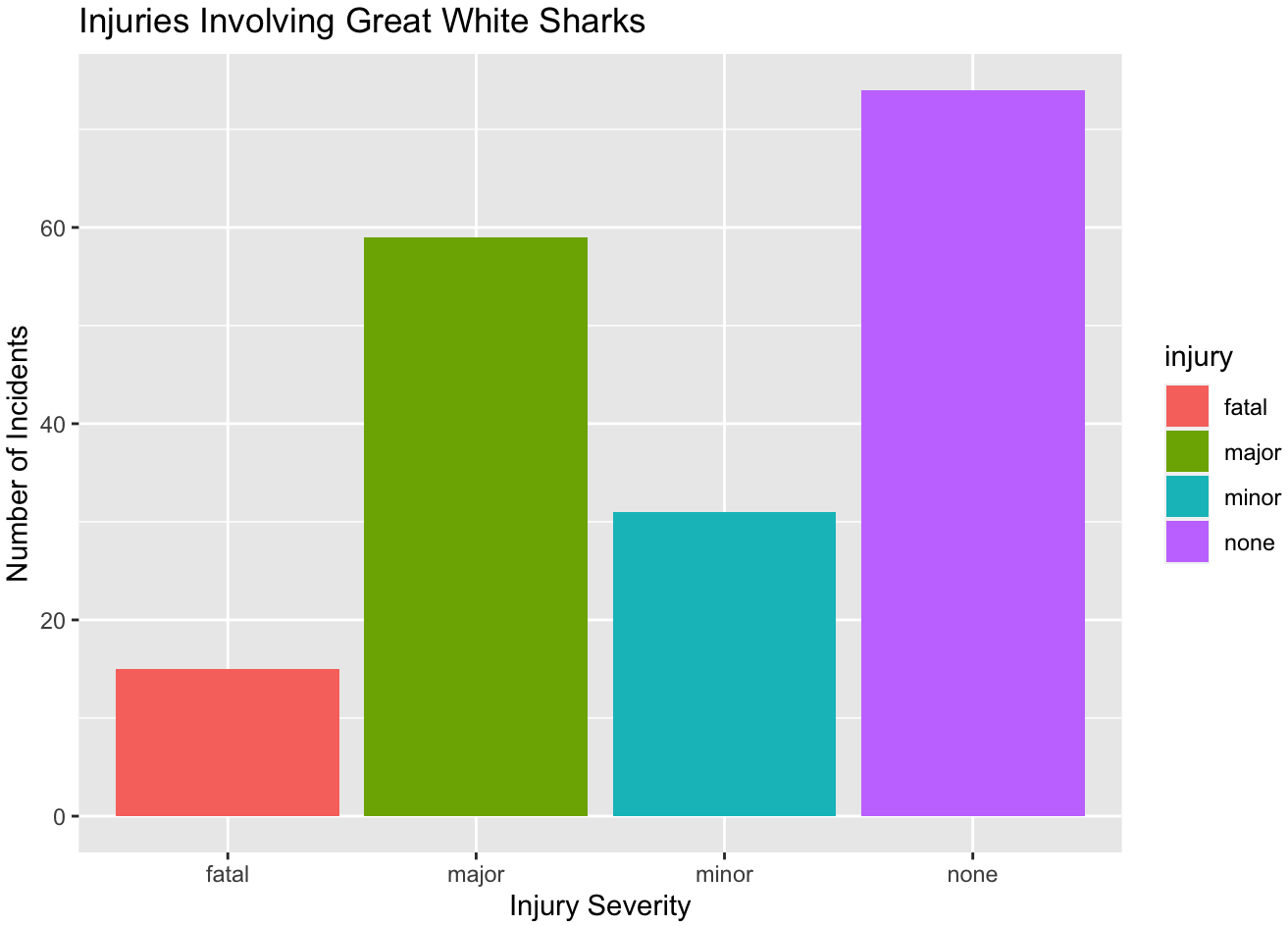
8. (1 point) Which shark species is involved in the highest number of incidents?

```
sharks %>%
  group_by(species) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  arrange(-number_incidents)
```

```
## # A tibble: 8 × 2
##   species    number_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
```

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") %>%
  ggplot(aes(injury, fill= injury)) +
  geom_bar() +
  labs(title = "Injuries Involving Great White Sharks",
       x= "Injury Severity",
       y= "Number of Incidents")
```



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?

```
head(white_sharks)
```

```
## # A tibble: 6 × 10
##   shark tagging_date total_length_cm sex   maturity   pop_up_date track_days
##   <chr> <chr>                  <dbl> <chr> <chr>      <chr>            <dbl>
## 1 1-M   19-Oct-99                402 M     Mature     2-Nov-99            14
## 2 2-M   30-Oct-99                366 M     Adolescent 25-Nov-99           26
## 3 3-M   16-Oct-00                457 M     Mature     16-Apr-01          182
## 4 4-M   5-Nov-01                 457 M     Mature     6-May-02           182
## 5 5-F   5-Nov-01                 488 F     Mature     19-Jul-02          256
## 6 6-M   5-Nov-01                 427 M     Mature     7-Aug-02           275
## # ℹ 3 more variables: longitude <dbl>, latitude <dbl>, comment <chr>
```

```
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…
```

```
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
```

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).

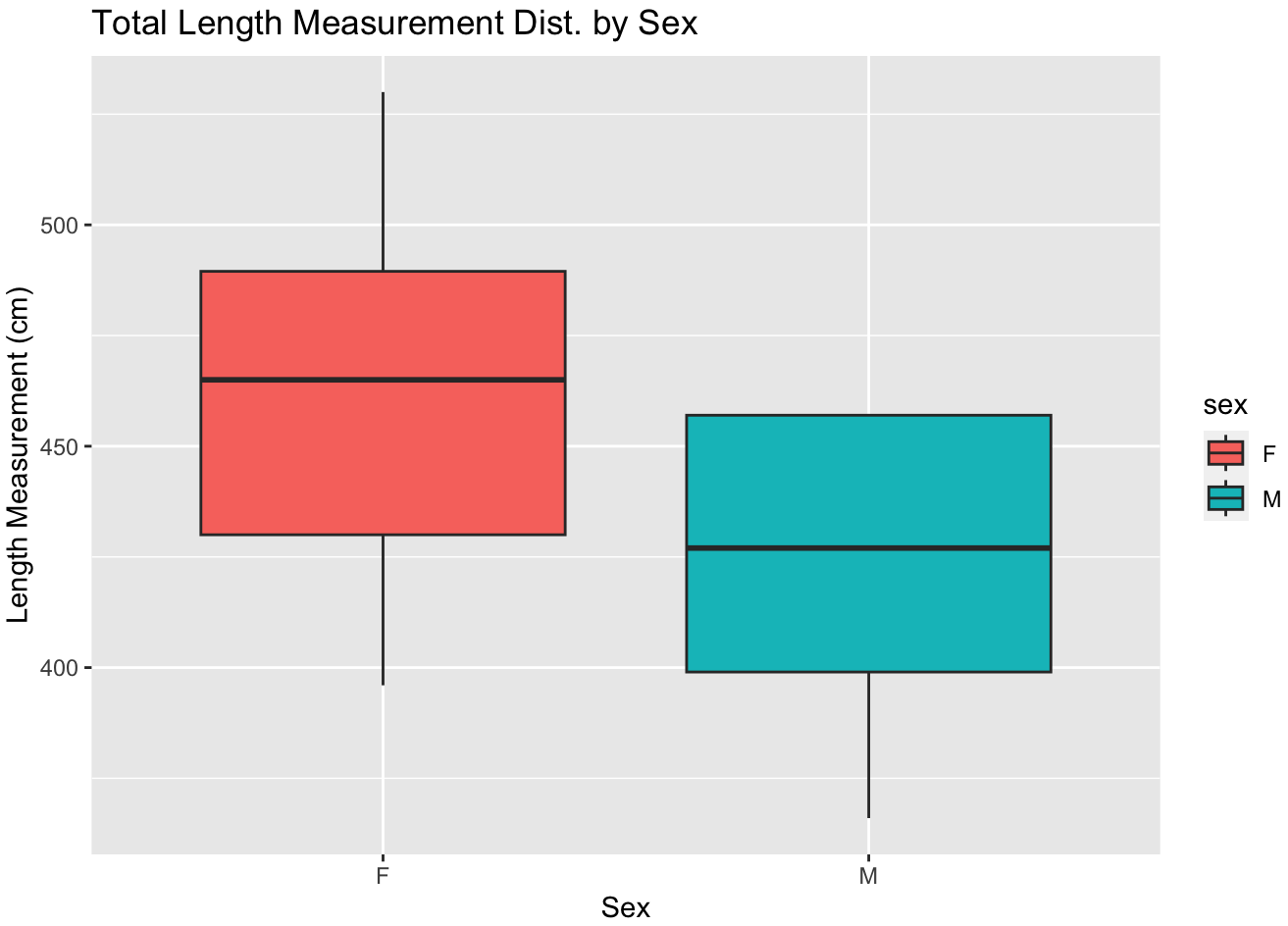
```
white_sharks %>%
  group_by(sex) %>%
  filter(sex != "NA") %>%
  summarise(mean_length = mean(total_length_cm))
```

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

```
Females in this sample are longer by approx. 0.5 meters. This finding is verified by a g
oogle search.
```

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

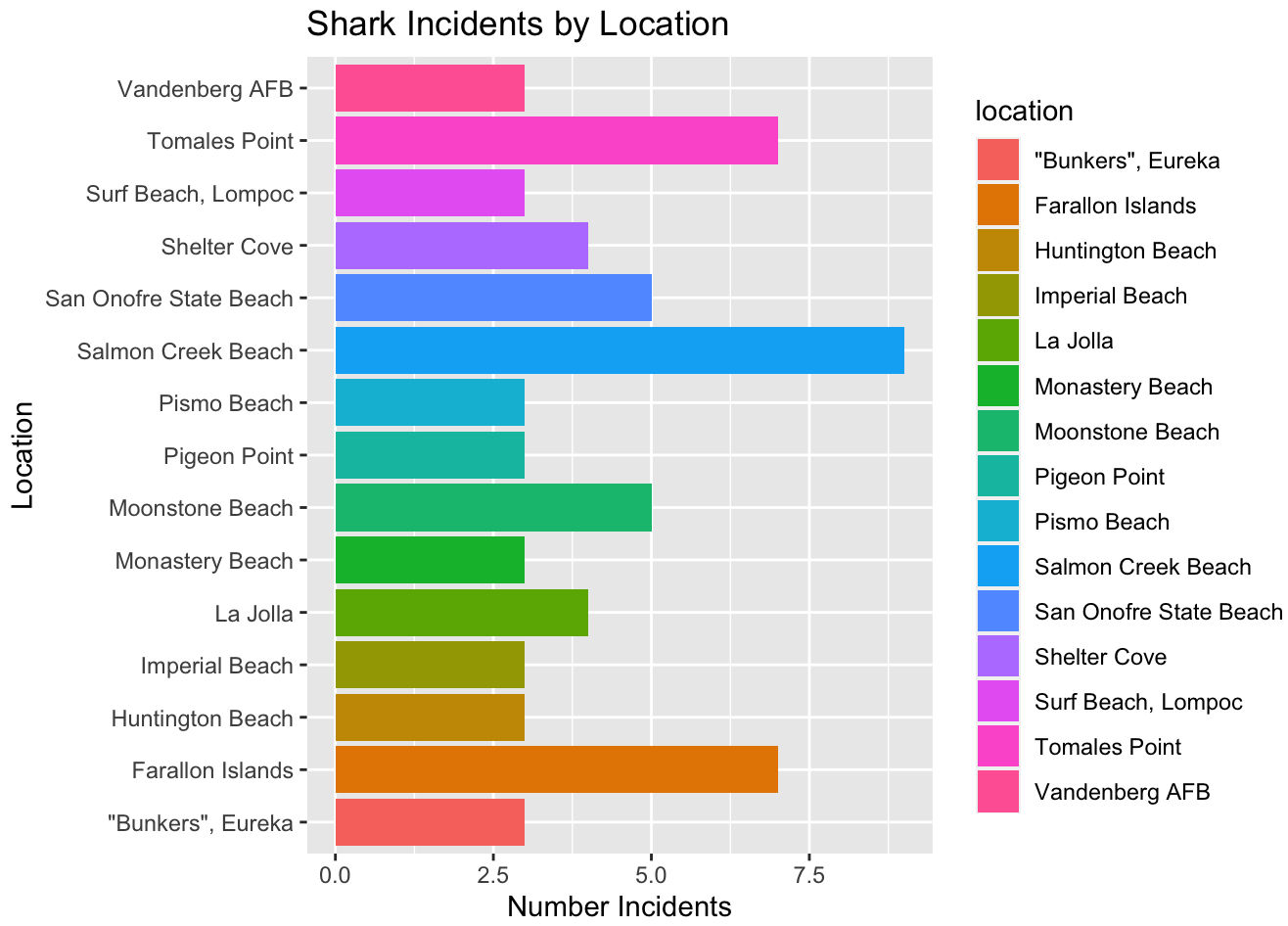
```
white_sharks %>%
  filter(sex != "NA") %>%
  ggplot(aes(x= sex, y=total_length_cm, fill= sex)) +
  geom_boxplot() +
  labs(title = "Total Length Measurement Dist. by Sex",
       x= "Sex",
       y= "Length Measurement (cm)")
```



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.

    I am going to answert the question: Which beach has the most recorded shark incidents

```
sharks %>%
  group_by(location) %>%
  summarise(number_incidents = n_distinct(incident_num)) %>%
  top_n(10, number_incidents) %>%
  arrange(-number_incidents) %>%
  ggplot(aes(x= location, y= number_incidents, fill= location)) +
  geom_col() +
  coord_flip() +
  labs(title = "Shark Incidents by Location",
       x= "Location",
       y= "Number Incidents")
```




```
Salmon Creek Beach is the location with the most incidents on this list.
```