Midterm 2 W24

Zhuoya Wang

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…
```

```
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
```

```
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>
```

```
sharks %>% naniar::miss_var_summary()
```

```
## # 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.

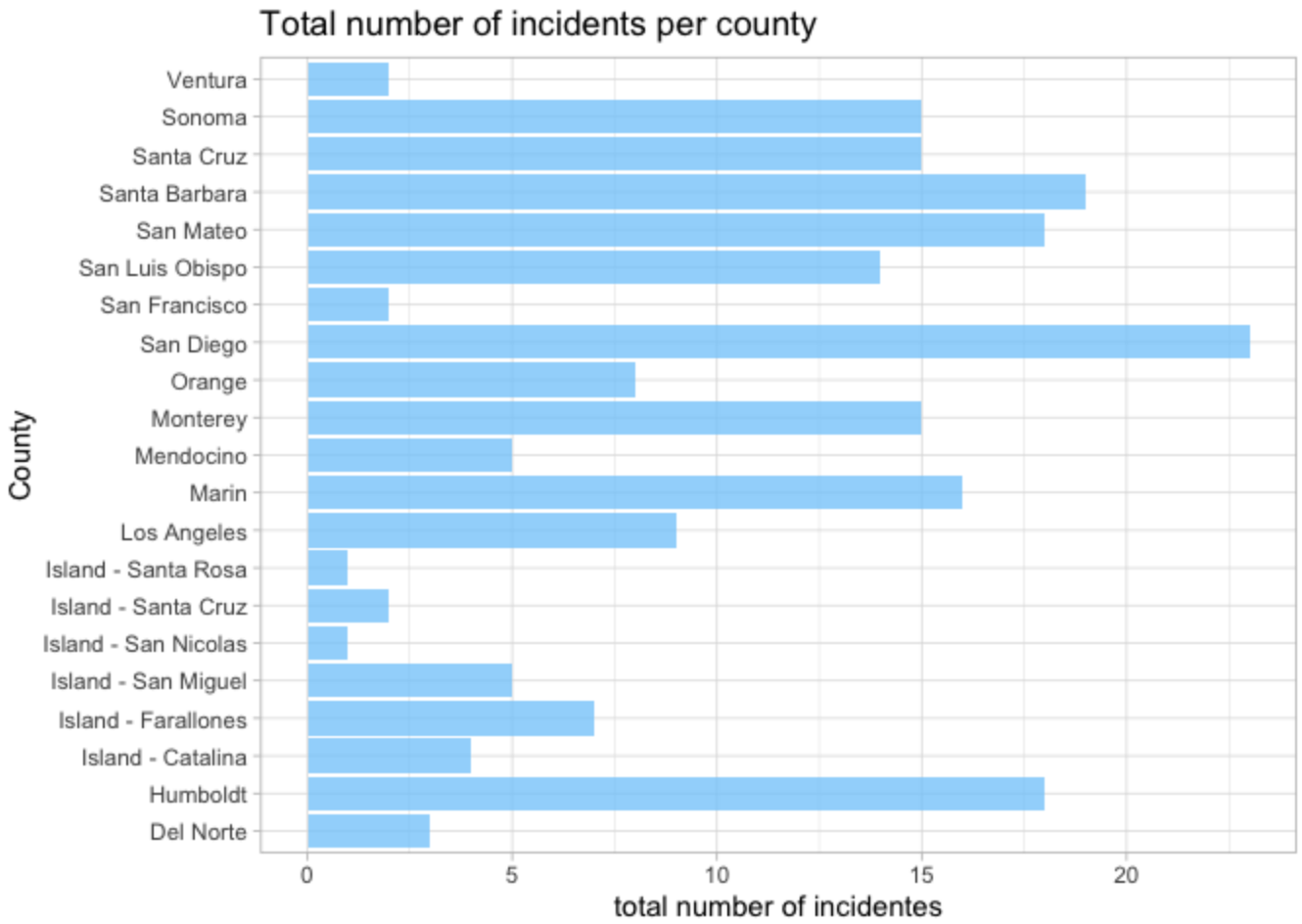
```
names(sharks)
```

```
##  [1] "incident_num"     "month"           "day"              "year"
##  [5] "time"             "county"          "location"         "mode"
##  [9] "injury"           "depth"           "species"          "comment"
## [13] "longitude"        "latitude"        "confirmed_source" "wfl_case_number"
```

```
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?

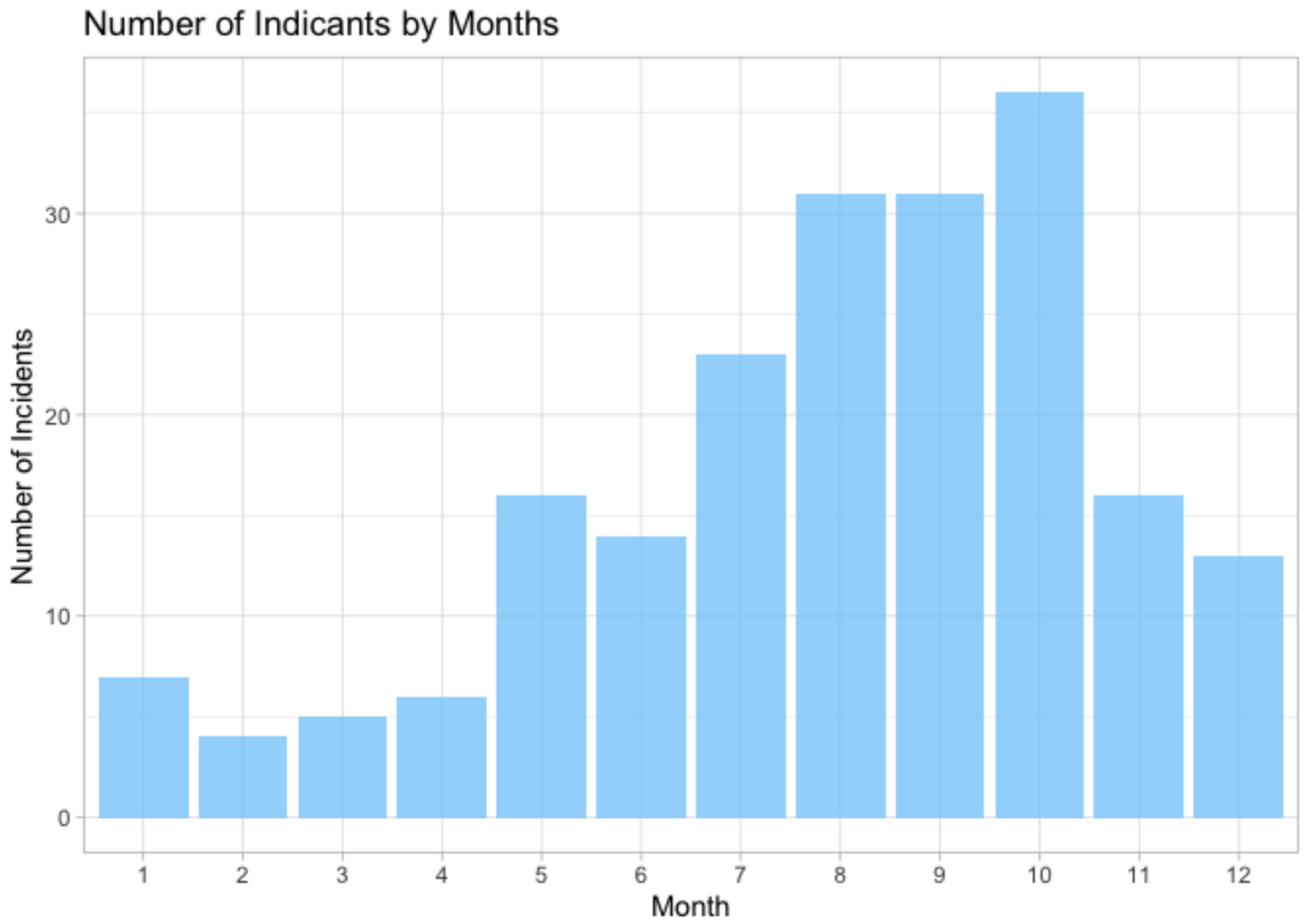
```
sharks%>%
  ggplot(aes(x = county))+
  geom_bar(fill = "lightskyblue", alpha = 0.8)+
  coord_flip()+
  labs(title = "Total number of incidents per county", y = "total number of incidentes",
x = "County")+
  theme_light()
```



San Diego has the highest number of incidents

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%>%
  ggplot(aes(x = as.factor(month)))+
  geom_bar(fill = "lightskyblue", alpha = 0.8)+
  labs(title = "Number of Indicants by Months", x="Month", y = "Number of Incidents")+
  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%>%
  tabyl(county, injury)
```

```
##                county fatal major minor none
##             Del Norte     0     0     2    1
##              Humboldt     0     7     2    9
##     Island - Catalina     0     0     1    3
##   Island - Farallones     0     7     0    0
##   Island - San Miguel     1     2     2    0
##  Island - San Nicolas     0     0     1    0
##   Island - Santa Cruz     0     0     1    1
##   Island - Santa Rosa     0     1     0    0
##           Los Angeles     1     0     6    2
##                 Marin     0     9     4    3
##             Mendocino     1     3     1    0
##              Monterey     2     8     2    3
##                Orange     0     1     2    5
##             San Diego     2     4     8    9
##         San Francisco     1     0     0    1
##       San Luis Obispo     3     3     1    7
##              San Mateo     1     1     4   12
##         Santa Barbara     2     2     6    9
##            Santa Cruz     1     3     3    8
##                Sonoma     0     8     1    6
##               Ventura     0     0     2    0
```

San Luis Obispo has the largest number of fatalities.

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

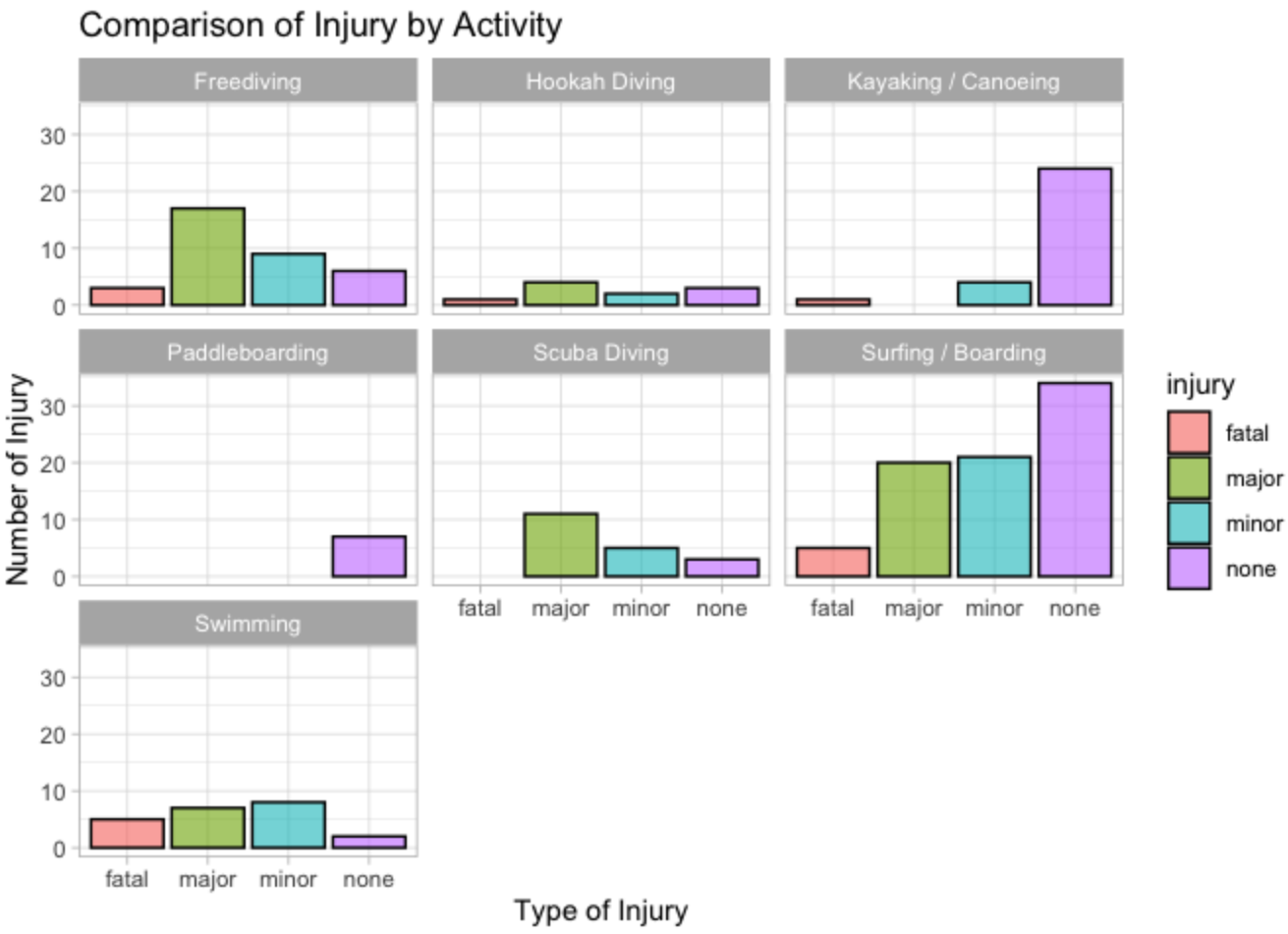
```
sharks%>%
  count(mode, sort = T)
```

```
## # A tibble: 7 × 2
##   mode                   n
##   <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
```

Surfing/Boarding is associated with 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%>%
  ggplot(aes(x=injury, fill=injury))+
  geom_bar( color="black", alpha=0.6, na.rm=T)+
  facet_wrap(.~ mode)+
  labs(title="Comparison of Injury by Activity", x="Type of Injury", y="Number of Injur
y")+
  theme_light()
```



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

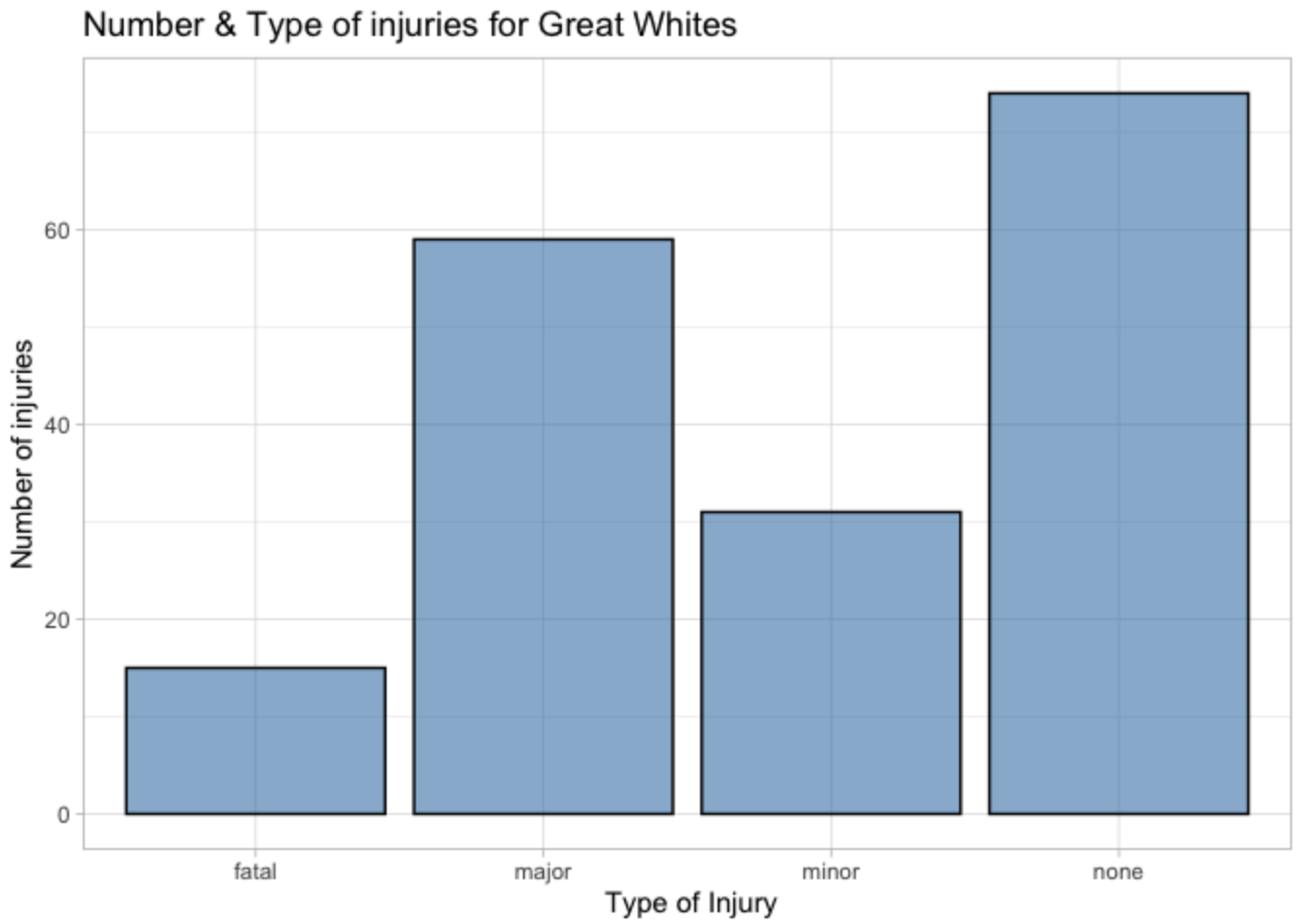
```
sharks%>%
  count(species, sort = T)
```

```
## # A tibble: 8 × 2
##   species        n
##   <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
```

White sharks are involved in the highest number of 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")%>%
  ggplot(aes(x=injury))+
  geom_bar(fill="steelblue", color="black", alpha=0.6, na.rm=T)+
  labs(title = "Number & Type of injuries for Great Whites", x = " Type of Injury",
       y = "Number of injuries")+
  theme_light()
```



Not all incidents involving Great White's 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 (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…
```

```
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
```

```
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>
```

```
white_sharks %>% naniar::miss_var_summary()
```

```
## # 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
```

```
names(white_sharks)
```

```
##  [1] "shark"           "tagging_date"    "total_length_cm" "sex"
##  [5] "maturity"        "pop_up_date"     "track_days"      "longitude"
##  [9] "latitude"        "comment"
```

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).

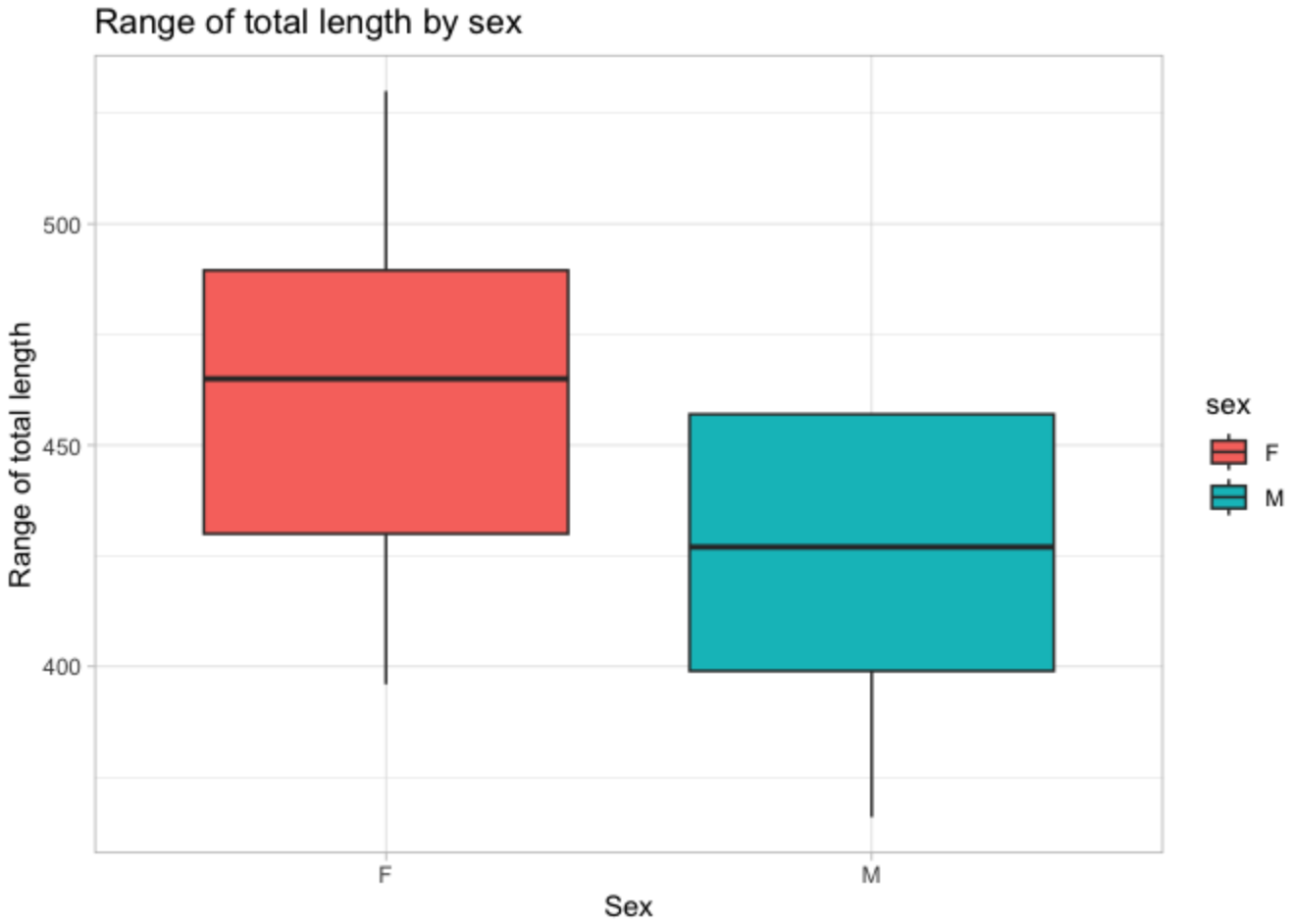
```
white_sharks%>%
  filter(sex != "NA")%>%
  group_by(sex)%>%
  summarize(mean_length = mean(total_length_cm))
```

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

In this data, Female white shark has a larger body length on average than male white sharks. I got the same results online (Females are generally bigger than males. Male great whites on average measure 3.4 to 4.0 m (11 to 13 ft) in length, while females measure 4.6 to 4.9 m (15 to 16 ft).)

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

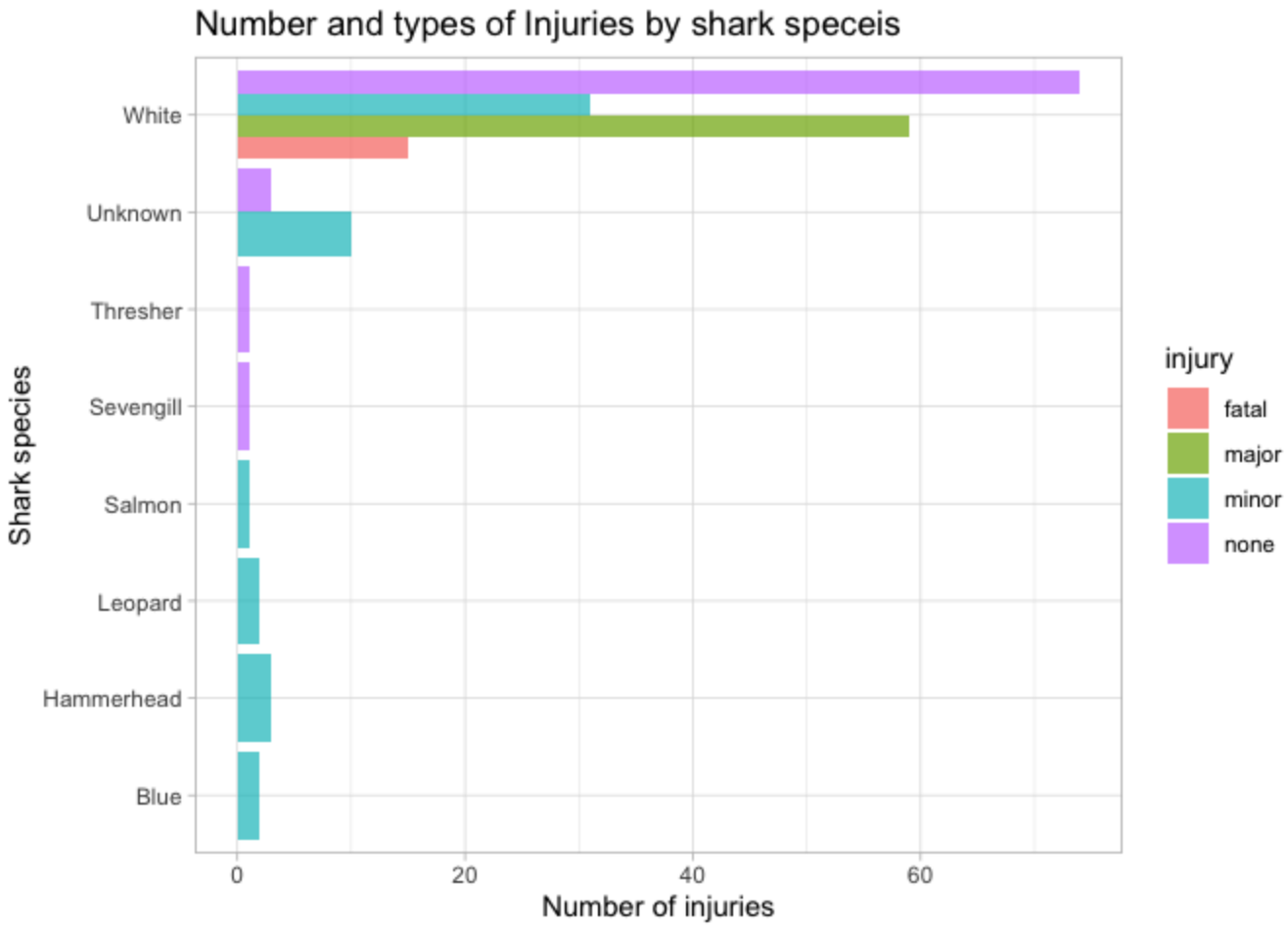
```
white_sharks%>%
  filter(sex != "NA")%>%
  group_by(sex)%>%
  ggplot(aes(x = sex, y = total_length_cm, fill = sex))+
  geom_boxplot()+
  labs(title = "Range of total length by sex", x = "Sex", y = "Range of total length")+
  theme_light()
```



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.

Question: make a plot that compares the number and types of injuries by shark speceis.

```
sharks %>%
  ggplot(aes(x=species, fill = injury)) +
  geom_bar(na.rm = T, position = "dodge", alpha=0.7)+
  coord_flip()+
  labs(title = "Number and types of Injuries by shark speceis", y = "Number of injurie
s", x = "Shark species")+
  theme_light()
```


Number and types of Injuries by shark speceis
White
Unknown
Thresher
Sevengill
Salmon
Leopard
Hammerhead
Blue
Shark species
0
20
40
60
Number of injuries
injury
fatal
major
minor
none