Midterm 2 W24

Carmen Doria

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?

The missing values are represented as NA’s and some appear to show up as unknown or not counted.

```
str(sharks)
```

```
## spc_tbl_ [211 × 16] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ incident_num    : chr [1:211] "1" "2" "3" "4" ...
##  $ month           : num [1:211] 10 5 12 2 8 4 10 5 6 7 ...
##  $ day             : num [1:211] 8 27 7 6 14 28 12 7 14 28 ...
##  $ year            : num [1:211] 1950 1952 1952 1955 1956 ...
##  $ time            : chr [1:211] "12:00" "14:00" "14:00" "12:00" ...
##  $ county          : chr [1:211] "San Diego" "San Diego" "Monterey" "Monterey" ...
##  $ location        : chr [1:211] "Imperial Beach" "Imperial Beach" "Lovers Point" "Pa
cific Grove" ...
##  $ mode            : chr [1:211] "Swimming" "Swimming" "Swimming" "Freediving" ...
##  $ injury          : chr [1:211] "major" "minor" "fatal" "minor" ...
##  $ depth           : chr [1:211] "surface" "surface" "surface" "surface" ...
##  $ species         : chr [1:211] "White" "White" "White" "White" ...
##  $ comment         : chr [1:211] "Body Surfing, bit multiple times on leg, thigh and
body" "Foot & swim fin bitten" "Attacked from below then second time from front, fatal"
"Attacked from behind, lost swim fin" ...
##  $ longitude       : chr [1:211] "-117.1466667" "-117.2466667" "-122.05" "-122.15"
...
##  $ latitude        : num [1:211] 32.6 32.6 36.6 36.6 35.1 ...
##  $ confirmed_source: chr [1:211] "Miller/Collier, Coronado Paper, Oceanside Paper" "G
SAF - with photos" "Miller/Collier, Coronado Paper" "Miller/Collier, Santa Cruz Sentine
l" ...
##  $ wfl_case_number : chr [1:211] NA NA NA NA ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   IncidentNum = col_character(),
##   ..   Month = col_double(),
##   ..   Day = col_double(),
##   ..   Year = col_double(),
##   ..   Time = col_character(),
##   ..   County = col_character(),
##   ..   Location = col_character(),
##   ..   Mode = col_character(),
##   ..   Injury = col_character(),
##   ..   Depth = col_character(),
##   ..   Species = col_character(),
##   ..   Comment = col_character(),
##   ..   Longitude = col_character(),
##   ..   Latitude = col_double(),
##   ..   `Confirmed Source` = col_character(),
##   ..   `WFL Case #` = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
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>
```

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.

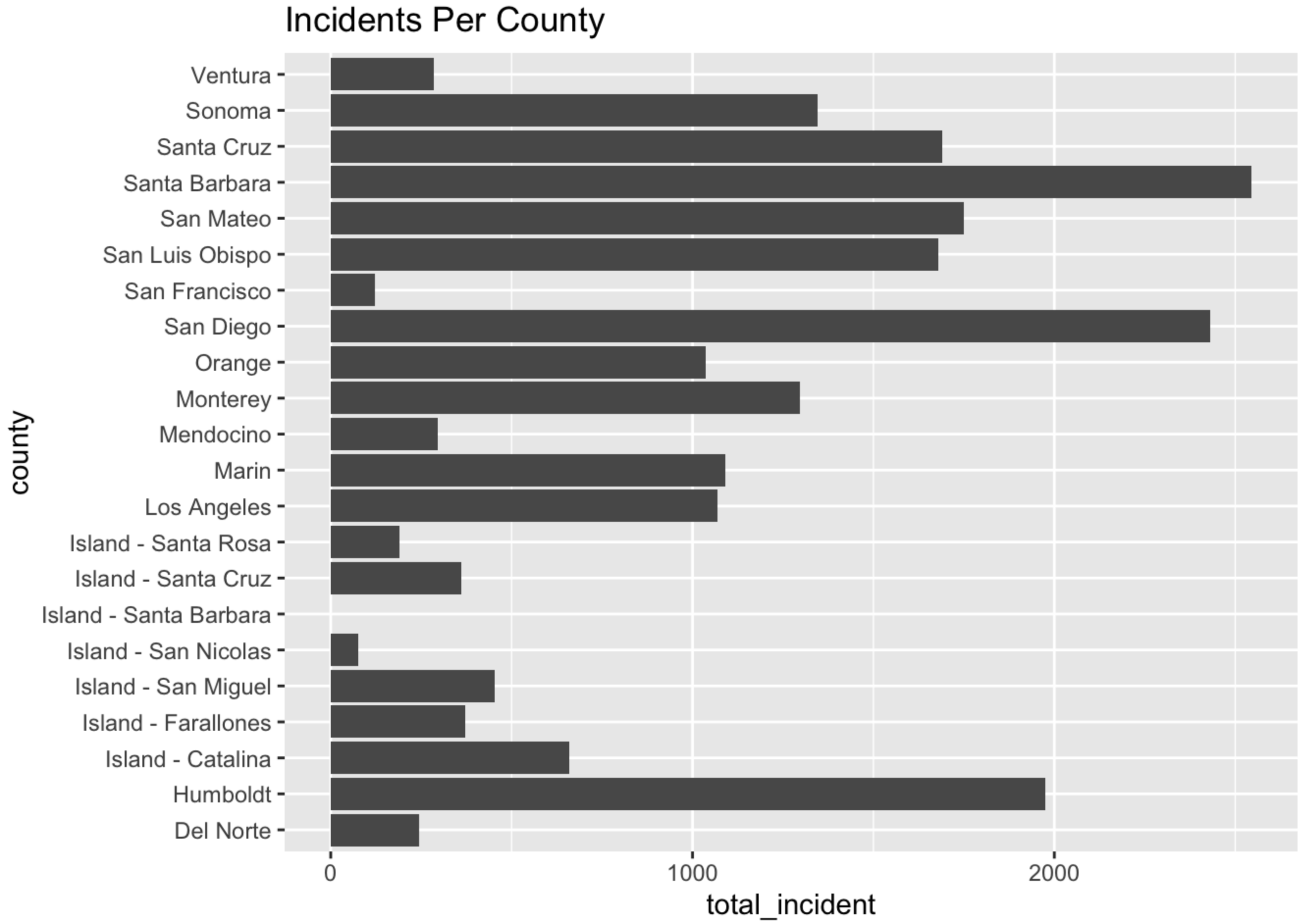
```
sharks <- sharks %>%
  replace_with_na_all(condition = ~.x == "NOT COUNTED")
```

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? It appears that Santa Barbara has the highest number of incidents.

```
names(sharks)
```

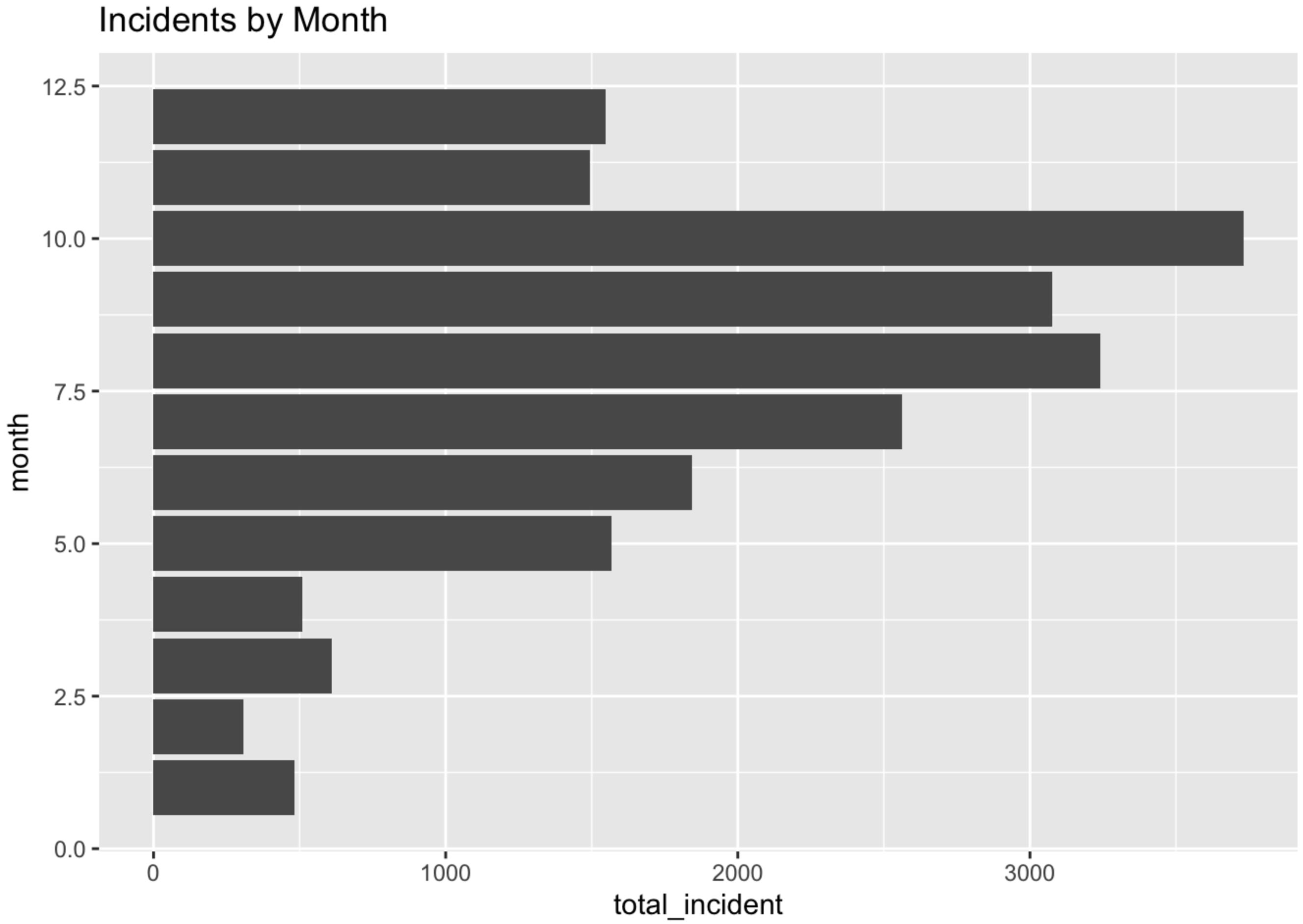
```
##  [1] "incident_num"     "month"            "day"              "year"
##  [5] "time"             "county"           "location"         "mode"
##  [9] "injury"           "depth"            "species"          "comment"
## [13] "longitude"        "latitude"         "confirmed_source" "wfl_case_number"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

```
sharks %>%
  select(county, incident_num) %>%
  group_by(county) %>%
  summarise(total_incident = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(total_incident)) %>%
  ggplot(aes(x = county, y = total_incident))+
  geom_col()+
  coord_flip()+
  labs(title = "Incidents Per County")
```



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? It appears that the most incidents occurred in October.

```
sharks %>%
  select(month, incident_num) %>%
  group_by(month) %>%
  summarise(total_incident = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(total_incident)) %>%
  ggplot(aes(x = month, y = total_incident))+
  geom_col()+
  coord_flip()+
  labs(title = "Incidents by Month")
```



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?
   San Luis Obispo has the highest number of fatalities.

```
sharks %>%
  select(mode, incident_num, county, injury) %>%
  filter(injury == "fatal") %>%
  group_by(county) %>%
  summarise(total_incident = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(total_incident))
```

```
## # A tibble: 10 × 2
##    county              total_incident
##    <chr>                        <int>
##  1 San Luis Obispo                310
##  2 Santa Barbara                  289
##  3 Santa Cruz                     192
##  4 San Diego                      136
##  5 Mendocino                      103
##  6 Island - San Miguel             82
##  7 Los Angeles                     62
##  8 San Mateo                       50
##  9 Monterey                        48
## 10 San Francisco                    8
```

```
sharks %>%
  select(mode, incident_num, county, injury) %>%
  group_by(county, injury) %>%
  summarise(total_incident = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(injury =="fatal"))
```

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

```
## # A tibble: 66 × 3
## # Groups:   county [22]
##    county              injury total_incident
##    <chr>               <chr>           <int>
##  1 Island - San Miguel fatal              82
##  2 Los Angeles         fatal              62
##  3 Mendocino           fatal             103
##  4 Monterey            fatal              48
##  5 San Diego           fatal             136
##  6 San Francisco       fatal               8
##  7 San Luis Obispo     fatal             310
##  8 San Mateo           fatal              50
##  9 Santa Barbara       fatal             289
## 10 Santa Cruz          fatal             192
## # ℹ 56 more rows
```

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

```
sharks %>%
  select(mode, incident_num) %>%
  group_by(mode) %>%
  summarise(total_incidents = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(total_incidents))
```

```
## # A tibble: 8 × 2
##   mode                total_incidents
##   <chr>                         <int>
## 1 Surfing / Boarding             9474
## 2 Kayaking / Canoeing            4317
## 3 Freediving                     2501
## 4 Swimming                       1816
## 5 Scuba Diving                   1332
## 6 Paddleboarding                  980
## 7 Hookah Diving                   550
## 8 Walking in shallow                0
```

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, y = mode))+
 # geom_bar()+
 # facet_wrap(~as.integer(incident_num))+
  #labs(title="Injuries by Activity", x="Injury Type", y=NULL)+
  #theme_light()
```

8. (1 point) Which shark species is involved in the highest number of incidents? It seems that the Great White sharks species are involved in the highest number of incidents.

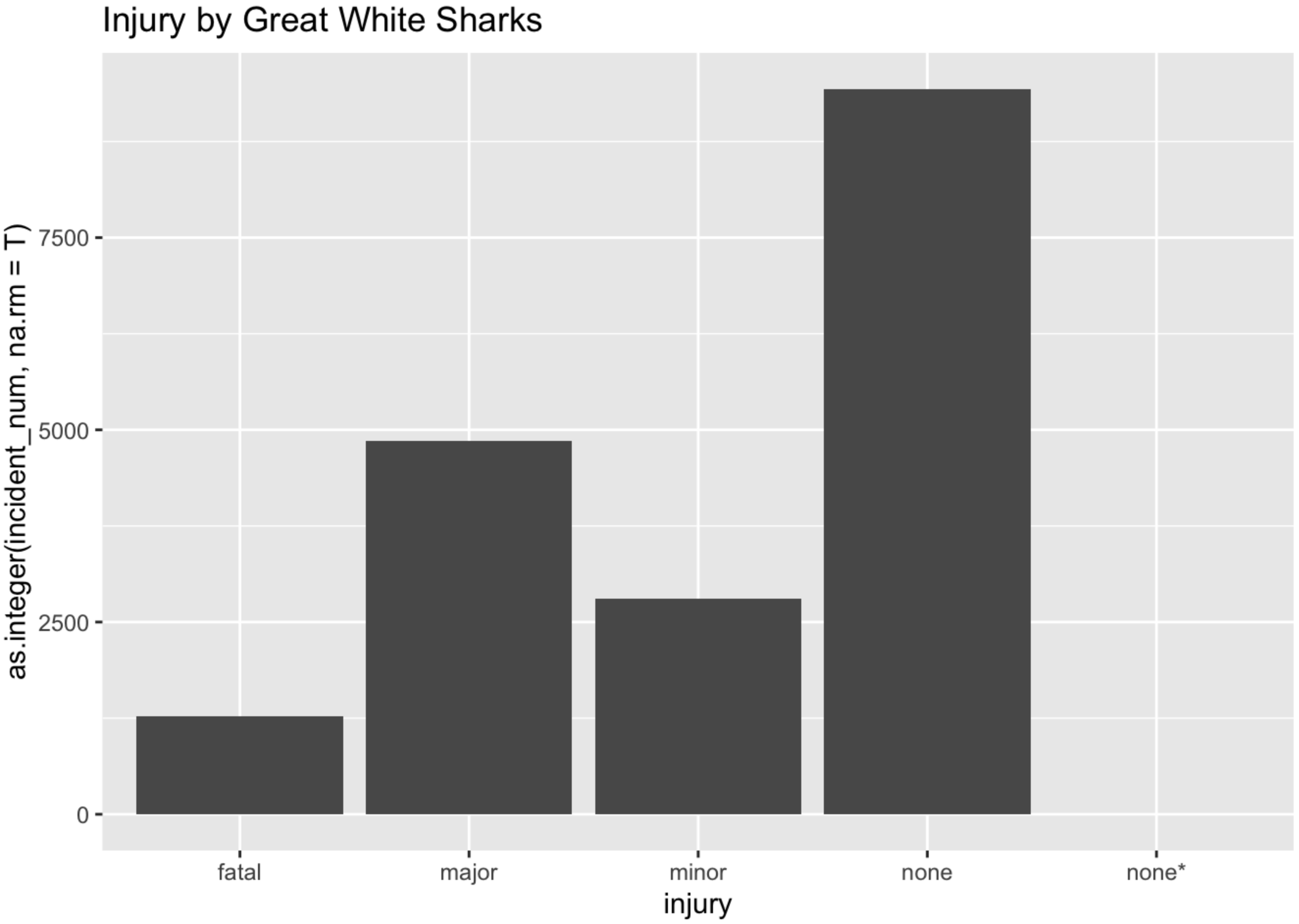
```
sharks %>%
  group_by(species) %>%
  summarise(total_incidents = sum(as.integer(incident_num), na.rm = T)) %>%
  arrange(desc(total_incidents))
```

```
## # A tibble: 12 × 2
##    species      total_incidents
##    <chr>                  <int>
##  1 White                  18374
##  2 Unknown                 1447
##  3 Hammerhead               344
##  4 Leopard                  261
##  5 Thresher                 187
##  6 Salmon                   154
##  7 Sevengill                140
##  8 Blue                      63
##  9 Blue*                      0
## 10 Killer Whale               0
## 11 Mako                       0
## 12 blue                       0
```

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 %>%
  select(injury, species, incident_num) %>%
  filter(species=="White") %>%
  ggplot(aes(x= injury, y= as.integer(incident_num, na.rm= T)))+
  geom_col()+
  labs(title = "Injury by Great White Sharks")
```

```
## Warning: Removed 2 rows containing missing values (`position_stack()`).
```



```
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"
```

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? The NAs are represented as NAs and they appear to be in the sex, longitude, latitude, and maturity columns.

```
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…
```

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).
    On average it appears that the females are larger.

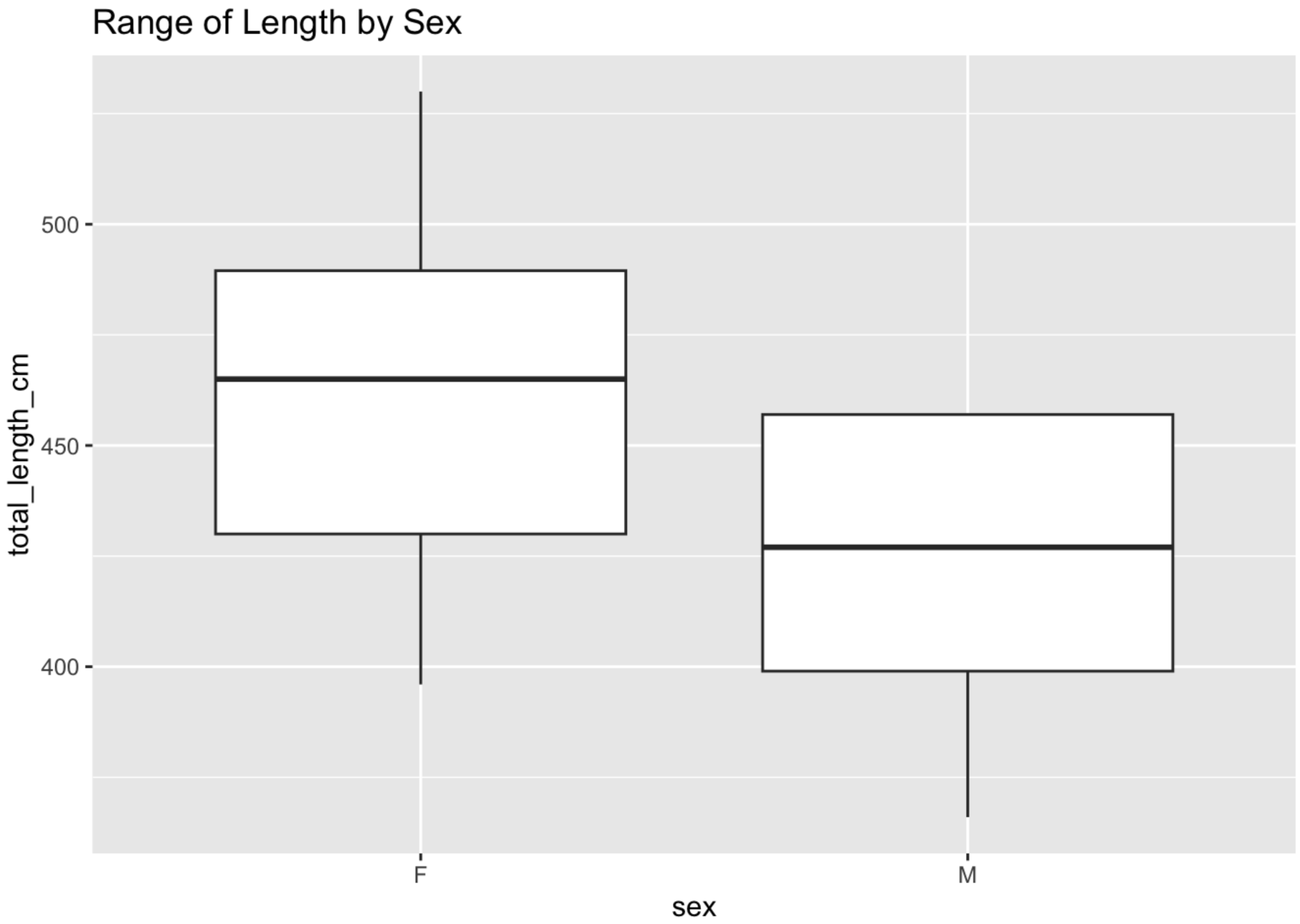
```
white_sharks %>%
  filter(sex == "F" | sex == "M") %>%
  group_by(sex) %>%
  summarise(mean_length = mean(total_length_cm))
```

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

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 = total_length_cm))+
  geom_boxplot(na.rm = T)+
  labs(title = "Range of Length by Sex")
```

```
## Warning: The following aesthetics were dropped during statistical transformation: fil
l
## ℹ This can happen when ggplot fails to infer the correct grouping structure in
##   the data.
## ℹ Did you forget to specify a `group` aesthetic or to convert a numerical
##   variable into a factor?
```



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 was the largest female whale found? Offshore focal area.

```
white_sharks %>%
  filter(sex == "F") %>%
  group_by(comment) %>%
  summarise(max_length = max(total_length_cm))
```

```
## # A tibble: 3 × 2
##   comment             max_length
##   <chr>                    <dbl>
## 1 Nearshore                  442
## 2 Offshore focal area        530
## 3 To Hawaii                  396
```