Midterm 2 W24

Amanda Duong

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 %>% 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
```

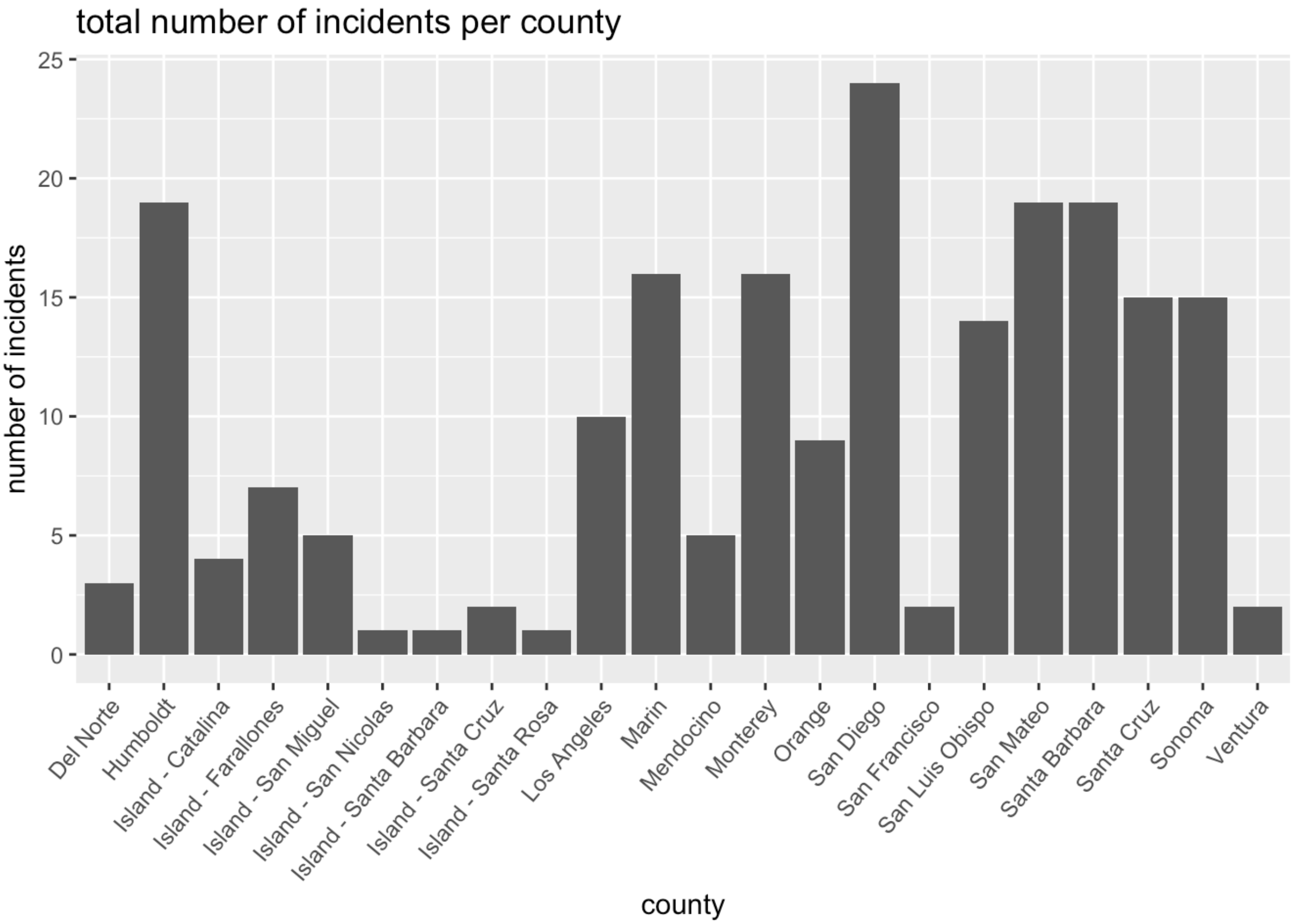
2. (1 point) Notice that there are some incidents identified as “NOT COUNTED”. These should be removed from the data because they were either not sharks, unverified, or were provoked. It’s OK to replace the `sharks` object.

```
sharks_new <- 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?

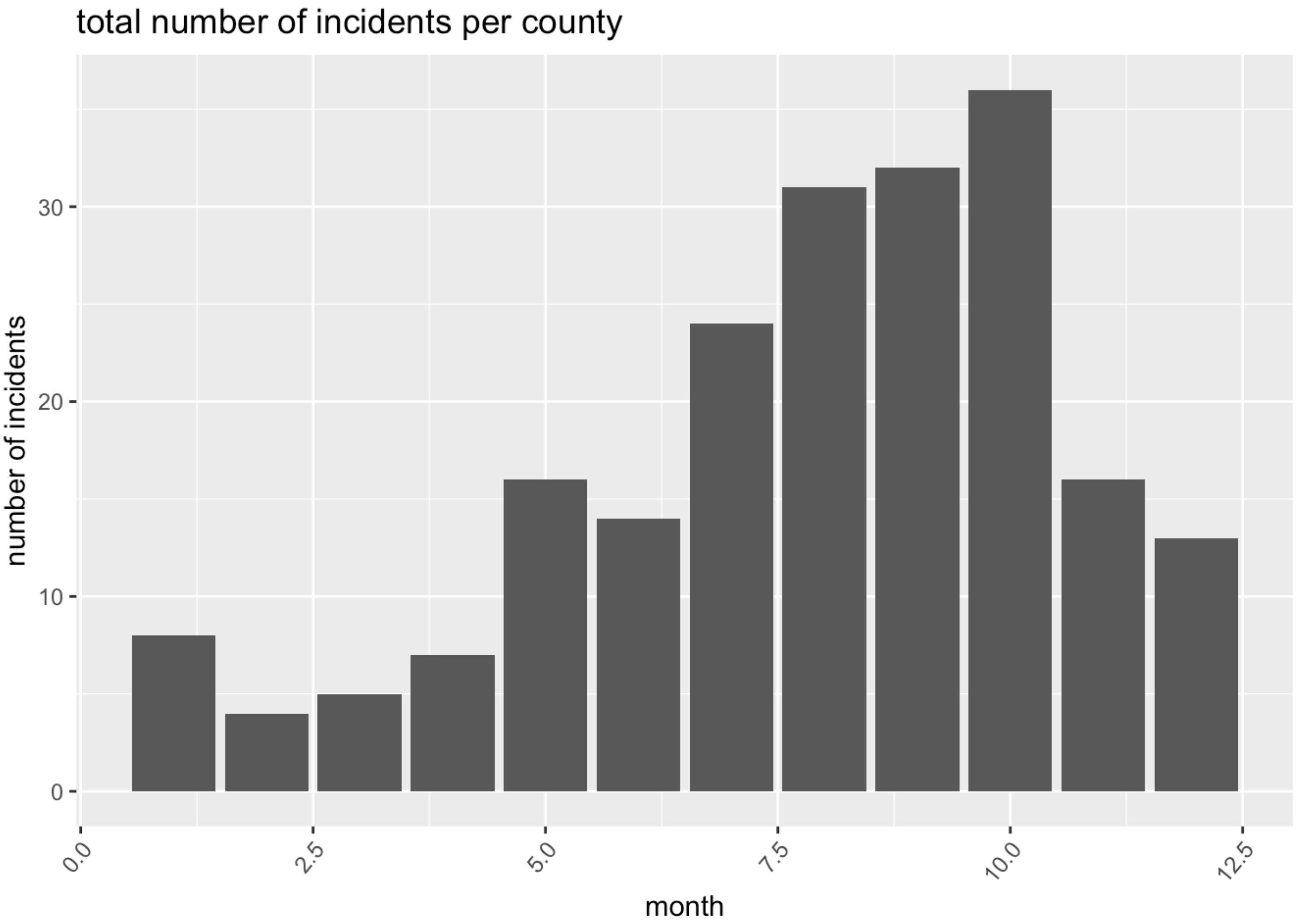
```
incidents <- sharks_new %>%
  group_by(county) %>%
  summarize(n_distinct_incident = n_distinct(incident_num))
#incidents
```

```
incidents %>%
  ggplot(aes(county, n_distinct_incident))+
  geom_col()+
  theme(axis.text.x=element_text(angle=50, hjust = 1))+
  labs(title="total number of incidents per county",
       y = "number of incidents",
       x= "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?

```
sharks_new %>%
  group_by(month) %>%
  summarize(n_distinct_incident = n_distinct(incident_num)) %>%
  ggplot(aes(month, n_distinct_incident))+
  geom_col()+
  theme(axis.text.x=element_text(angle=50, hjust = 1))+
  labs(title="total number of incidents per county",
       y = "number of incidents",
       x= "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?

```
sharks_new %>%
  select(injury, county, incident_num) %>%
  group_by(injury, county) %>%
  mutate(n_distinct_incident = n_distinct(incident_num))
```

```
## # A tibble: 211 × 4
## # Groups:   injury, county [66]
##    injury county          incident_num n_distinct_incident
##    <chr>  <chr>           <chr>                      <int>
##  1 major  San Diego       1                              4
##  2 minor  San Diego       2                              8
##  3 fatal  Monterey        3                              2
##  4 minor  Monterey        4                              2
##  5 major  San Luis Obispo 5                              3
##  6 fatal  San Luis Obispo 6                              3
##  7 major  San Diego       7                              4
##  8 fatal  San Francisco   8                              1
##  9 fatal  San Diego       9                              2
## 10 minor  San Diego       10                             8
## # ℹ 201 more rows
```

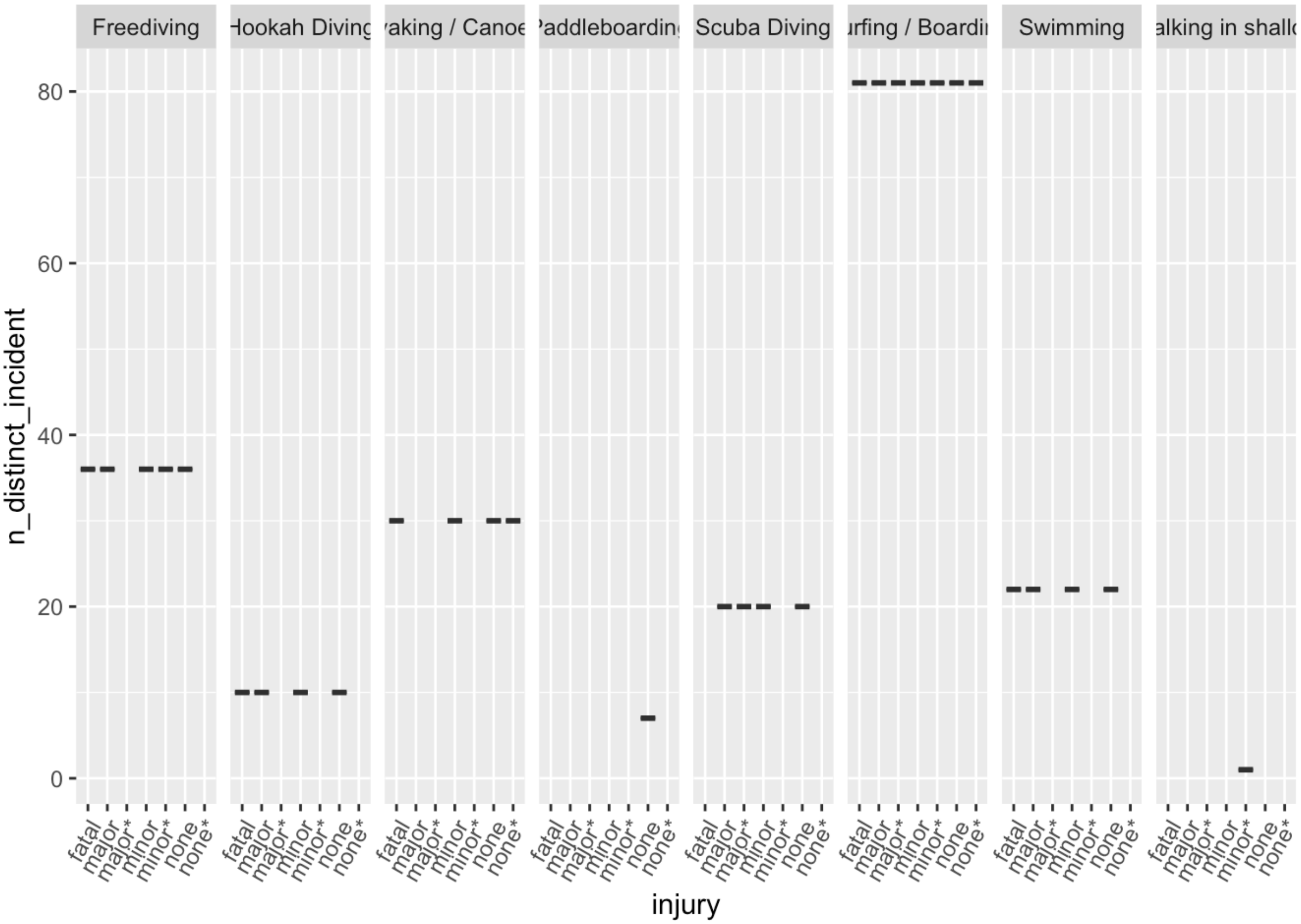
Most fatalities is island San Miguel. 6. (2 points) In the data, `mode` refers to a type of activity. Which activity is associated with the highest number of incidents?

```
mode_incident <- sharks_new %>%
  select(incident_num, injury, mode) %>%
  group_by(mode) %>%
  summarize(n_distinct_incident = n_distinct(incident_num)) %>%
  arrange(desc(n_distinct_incident))
#mode_incident
```

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_new %>%
  group_by(mode) %>%
  mutate(n_distinct_incident = n_distinct(incident_num)) %>%
  ggplot(aes(injury, n_distinct_incident))+
  geom_boxplot()+
  facet_grid(.~mode)+
  theme(axis.text.x = element_text(angle = 60, hjust=1))
```



8.

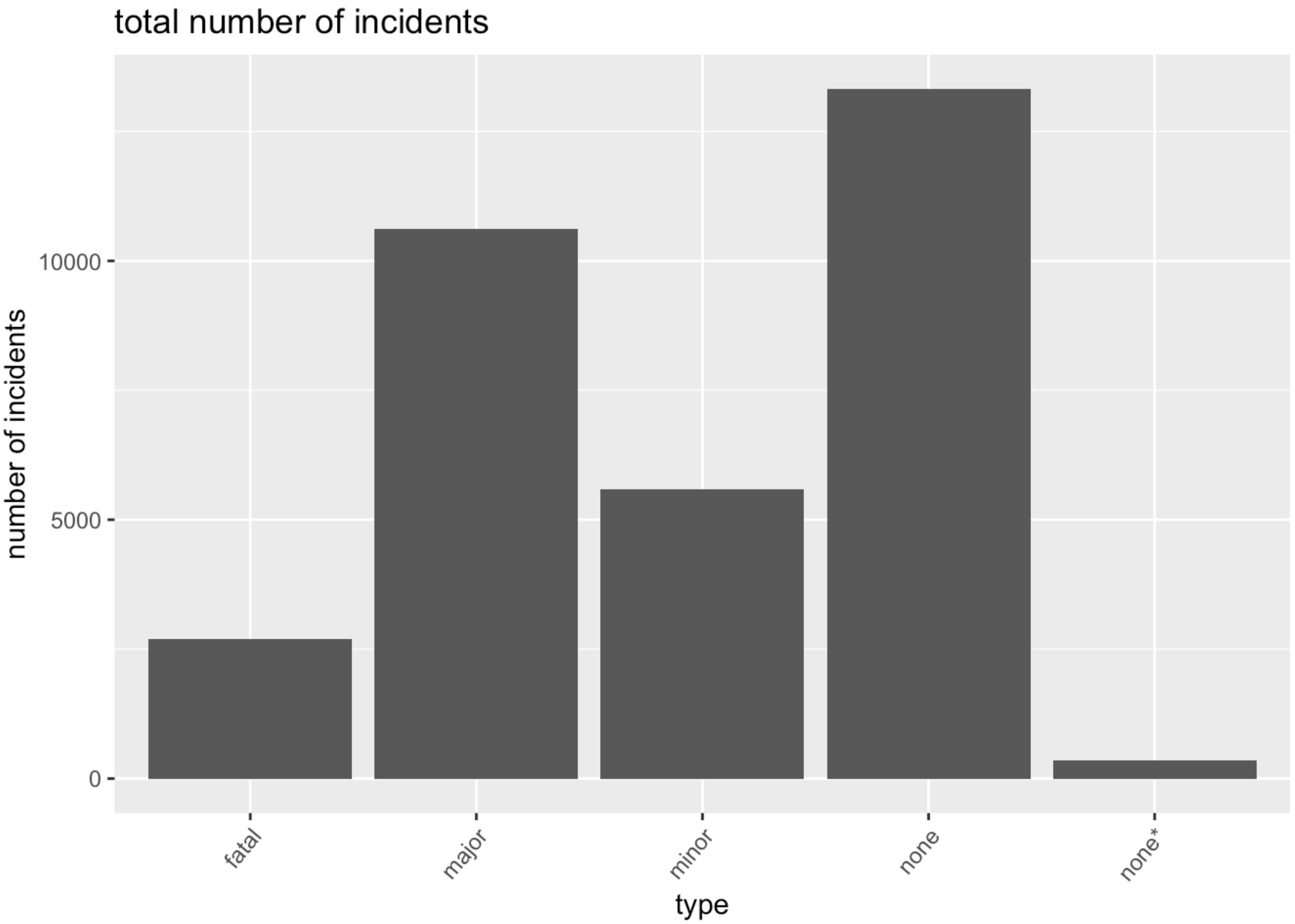
(1 point) Which shark species is involved in the highest number of incidents?

```
sharks_new %>%
  group_by(species) %>%
  summarize(n_distinct_incident = n_distinct(incident_num)) %>%
  arrange(desc(n_distinct_incident))
```

```
## # A tibble: 12 × 2
##    species      n_distinct_incident
##    <chr>                      <int>
##  1 White                        180
##  2 Unknown                       14
##  3 Hammerhead                     3
##  4 Blue                           2
##  5 Leopard                        2
##  6 Salmon                         2
##  7 Blue*                          1
##  8 Killer Whale                   1
##  9 Mako                           1
## 10 Sevengill                      1
## 11 Thresher                       1
## 12 blue                           1
```

Great White Shark is involved in the most 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_new %>%
  group_by(species) %>%
  mutate(n_distinct_incident = n_distinct(incident_num)) %>%
  filter(species == "White") %>%
  ggplot(aes(injury, n_distinct_incident))+
  geom_col()+
  theme(axis.text.x=element_text(angle=50, hjust = 1))+
  labs(title="total number of incidents",
       y = "number of incidents",
       x= "type")
```



no, most are major injuries.

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, C
A, 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 %>% 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
```

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 %>%
  select(sex, total_length_cm) %>%
  group_by(sex) %>%
  summarise(total_length = sum(total_length_cm, na.rm=T)) %>%
  arrange(desc(total_length))
```

```
## # A tibble: 3 × 2
##   sex   total_length
##   <chr>        <dbl>
## 1 M             4676
## 2 F             2772
## 3 <NA>          1275
```

Males are on average larger. 12. (3 points) Make a plot that compares the range of total length by sex.

```
#white_sharks %>%
  #group_by(sex) %>%
  #mutate(total_length = sum(total_length_cm, na.rm=T)) %>%
  #ggplot(aes(sex, total_length))+
  #geom_bar()+
  #labs(title = "Total Length by Sex",
      # x="sex",
      # y="total length")
```

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.