Homework 11

Sameeksha Deshatty

2025-02-20

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 final lab report should be organized, clean, and run free from errors. Remember, you must remove the # for the included code chunks to run. Be sure to add your name to the author header above.

Make sure to use the formatting conventions of RMarkdown to make your report neat and clean!

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. 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
```

```
miss_var_summary(sharks)
```

```
## # A tibble: 16 x 3
##    variable         n_miss pct_miss
##    <chr>             <int>    <num>
##  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. 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 %>%
  mutate(incident_num = as.numeric(incident_num)) %>%

filter(!is.na(incident_num))
```

```
## Warning: There was 1 warning in `mutate()`.
## i In argument: `incident_num = as.numeric(incident_num)`.
## Caused by warning:
## ! NAs introduced by coercion
```

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

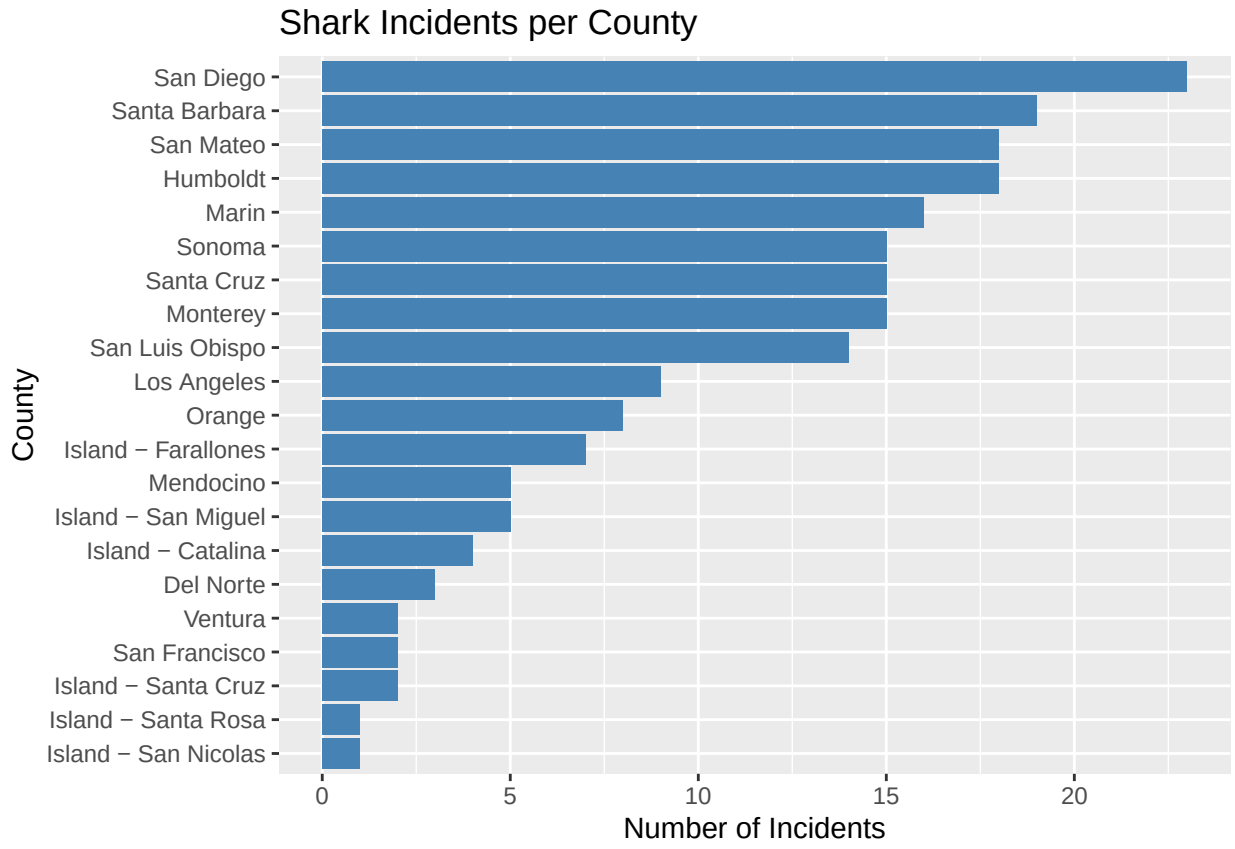
```
county_counts <- sharks %>%
  count(county, sort = TRUE)

print(county_counts)
```

```
## # A tibble: 21 x 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
```

```
ggplot(county_counts, aes(x = reorder(county, n), y = n)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  coord_flip() +
  labs(title = "Shark Incidents per County", x = "County", y = "Number of Incidents")
```



San Diego has the highest number of incidents.

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

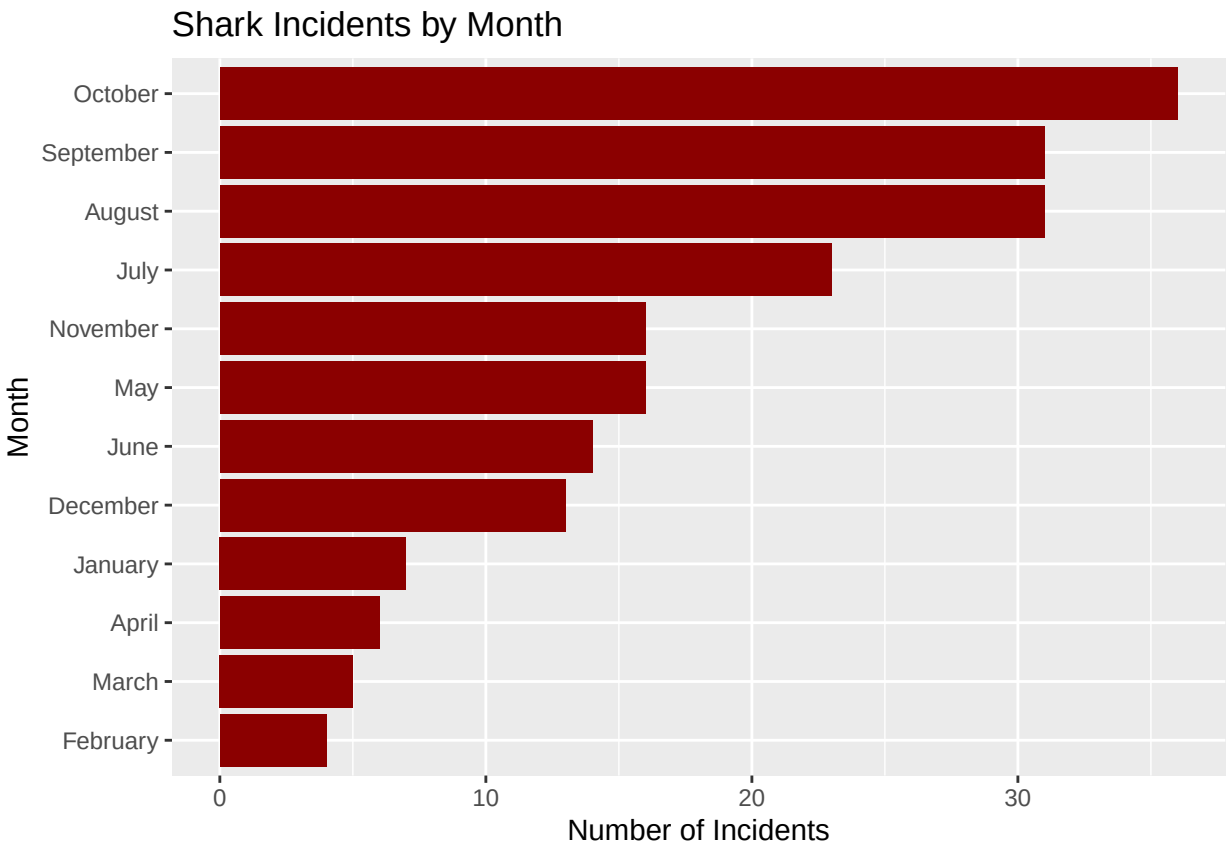
```
sharks <- sharks %>%
  mutate(date = as.Date(paste(year, month, day, sep = "-"), format = "%Y-%m-%d"))

sharks <- sharks %>%
  mutate(month_name = format(date, "%B"))

month_counts <- sharks %>%
  count(month_name, sort = TRUE)

ggplot(month_counts, aes(x = reorder(month_name, n), y = n)) +
  geom_bar(stat = "identity", fill = "darkred") +
```

```
  coord_flip() +
  labs(title = "Shark Incidents by Month", x = "Month", y = "Number of Incidents")
```



Incidents are most likely to occur in October.

5. How do the number and types of injuries compare by county? Make a table that shows the number of injury types by county. Which county has the highest number incidents?

```
injury_by_county <- sharks %>%
  count(county, injury, sort = TRUE)

print(injury_by_county)
```

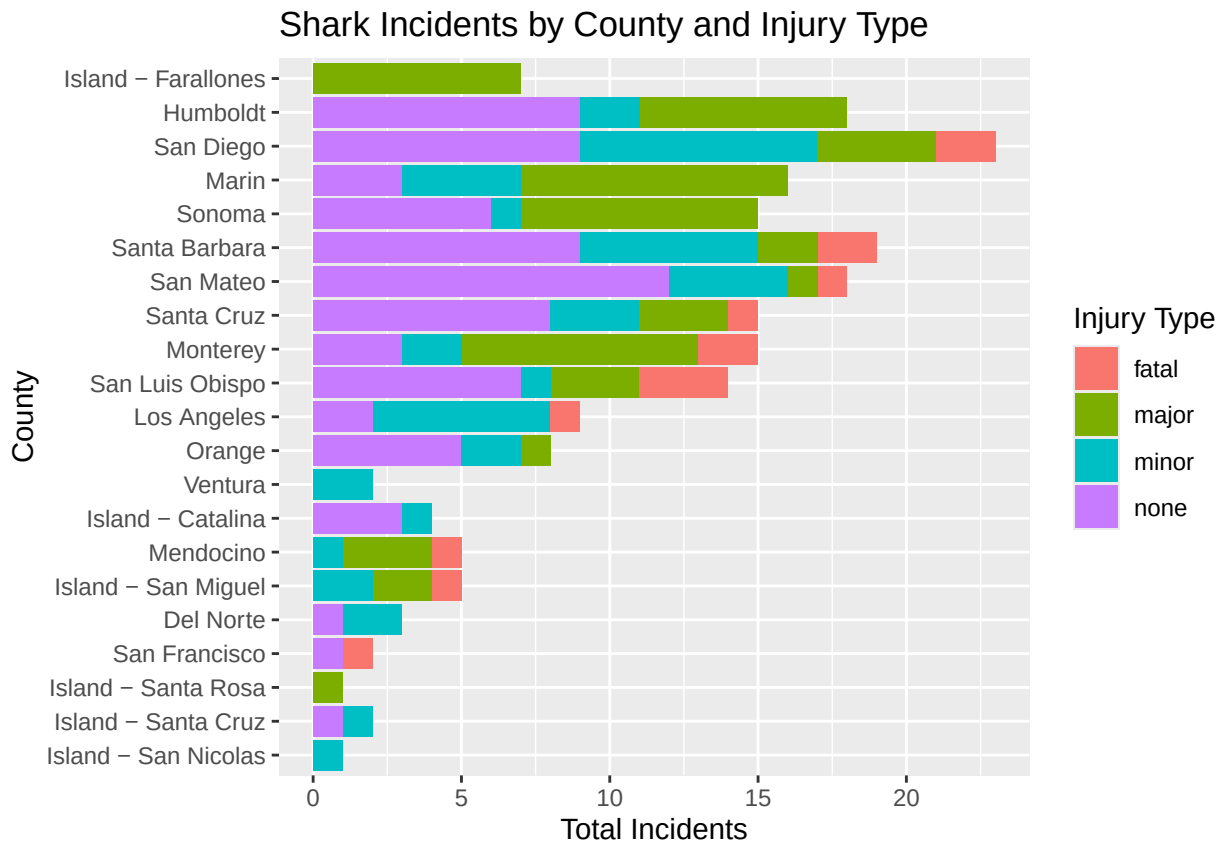
```
## # A tibble: 57 x 3
##    county        injury     n
##    <chr>         <chr>  <int>
##  1 San Mateo     none      12
##  2 Humboldt      none       9
##  3 Marin         major      9
##  4 San Diego     none       9
##  5 Santa Barbara none       9
##  6 Monterey      major      8
##  7 San Diego     minor      8
##  8 Santa Cruz    none       8
##  9 Sonoma        major      8
```

```
## 10 Humboldt      major      7
## # i 47 more rows
```

San Mateo has the highest number of incidents.

6. Use the table from #5 to make a plot that shows the total number of incidents by county.

```
ggplot(injury_by_county, aes(x = reorder(county, n), y = n, fill = injury)) +
  geom_bar(stat = "identity") +
  coord_flip() +
  labs(title = "Shark Incidents by County and Injury Type", x = "County", y = "Total Incidents", fill =
```



7. In the data, `mode` refers to a type of activity. Which activity is associated with the highest number of incidents?

```
activity_counts <- sharks %>%
  count(mode, sort = TRUE)

print(activity_counts)
```

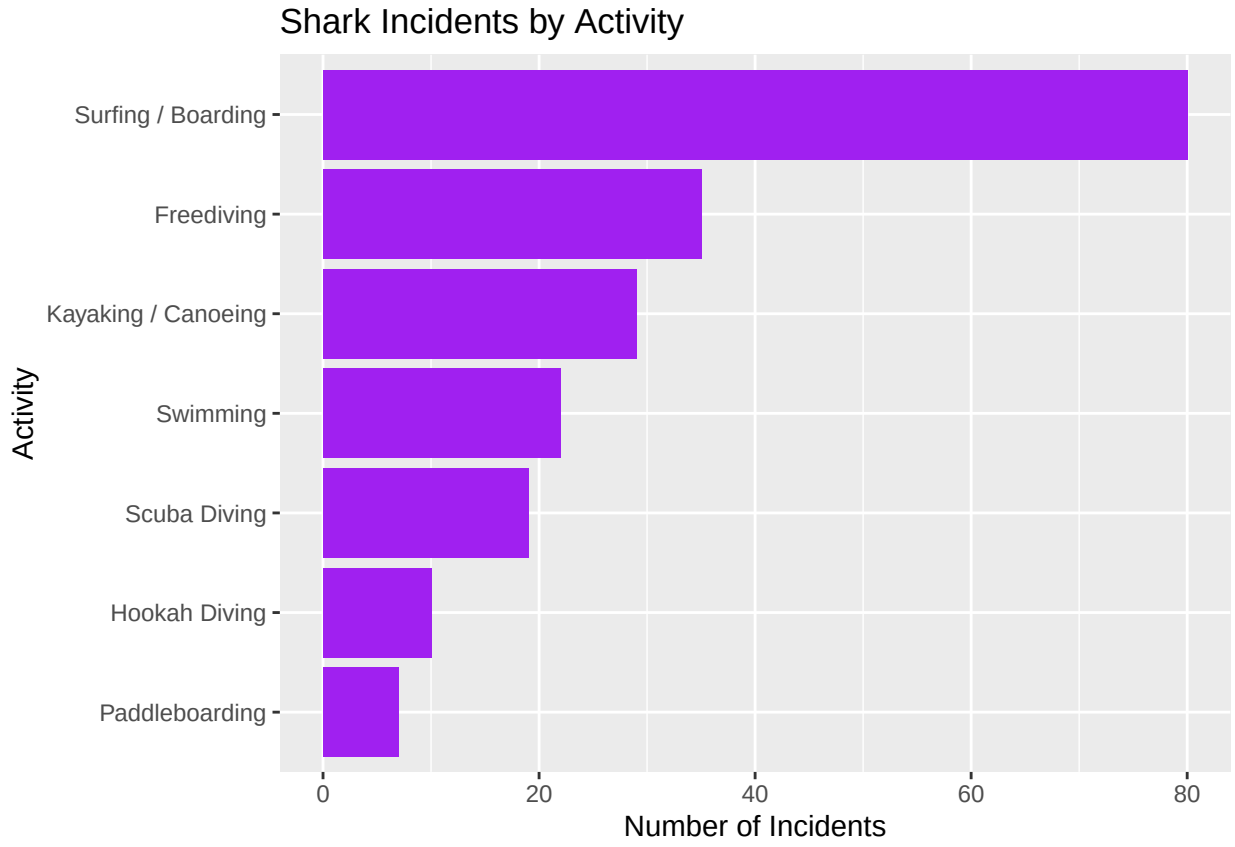
```
## # A tibble: 7 x 2
##   mode                  n
##   <chr>             <int>
## 1 Surfing / Boarding    80
```

```
## 2 Freediving             35
## 3 Kayaking / Canoeing    29
## 4 Swimming               22
## 5 Scuba Diving           19
## 6 Hookah Diving          10
## 7 Paddleboarding          7
```

Surfing is involved in the highest number of incidents.

8. Make a plot that compares the number of incidents by activity.

```
ggplot(activity_counts, aes(x = reorder(mode, n), y = n)) +
  geom_bar(stat = "identity", fill = "purple") +
  coord_flip() +
  labs(title = "Shark Incidents by Activity", x = "Activity", y = "Number of Incidents")
```



9. Which shark species is involved in the highest number of incidents?

```
species_counts <- sharks %>%
  count(species, sort = TRUE)

print(species_counts)
```

```
## # A tibble: 8 x 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
```

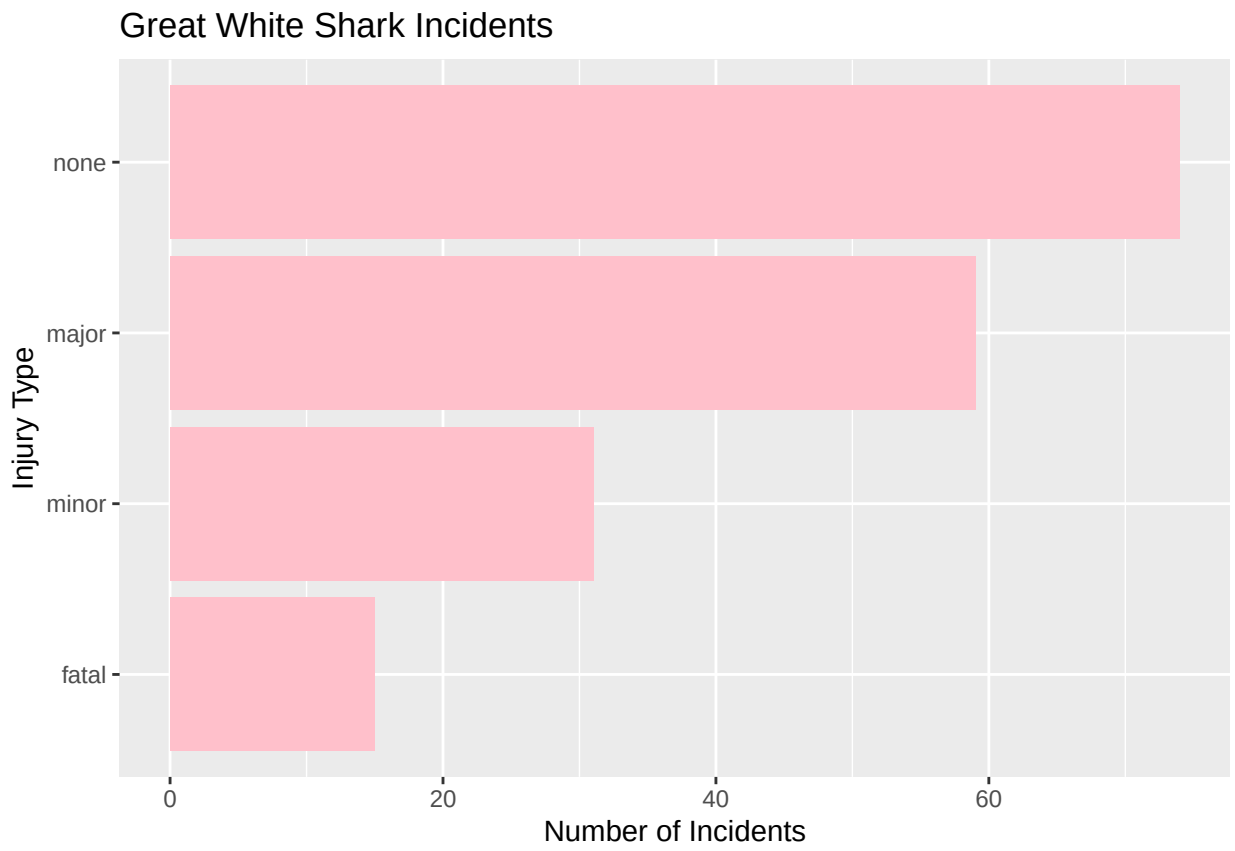
Great White Sharks are the most involved in accidents.

10. Are all incidents involving Great White's fatal? Make a plot that shows the number and types of incidents for Great White's only.

```
great_white_incidents <- sharks %>%
  filter(species == "White")

great_white_summary <- great_white_incidents %>%
  count(injury, sort = TRUE)

ggplot(great_white_summary, aes(x = reorder(injury, n), y = n)) +
  geom_bar(stat = "identity", fill = "pink") +
  coord_flip() +
  labs(title = "Great White Shark Incidents", x = "Injury Type", y = "Number of Incidents")
```



Not all incidents involving Great White Sharks are fatal.