````
---
title: "Midterm 2 W24"
author: "Maya Medina"
date: "`r Sys.Date()`"
output:
  html_document: default
  pdf_document: default
---

## 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
```{r message=FALSE, warning=FALSE}
library("tidyverse")
library("janitor")
library("naniar")
```

## Load the data
Run the following code chunk to import the data.
```{r message=FALSE, warning=FALSE}
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?
```{r}
str(sharks)
```

```{r}
any_na(sharks)
```
The missing values are marked as NAs.
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.
```{r}
sharks <- sharks %>%
  mutate(incident_num=na_if(incident_num, "NOT COUNTED")) %>%
````
````

```
   filter(!is.na(incident_num))
sharks
```
```{r}
sharks %>%
  mutate(incident_num, )
```

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?
```{r}
sharks %>%
  select(incident_num, county) %>%
  group_by(county) %>%
  count(county) %>%
  ggplot(aes(x=county, y=n))+
  geom_point(na.rm=T)+
  theme(axis.text.x=element_text(angle=40, hjust=1))+
  labs(title="Shark Incidents by County",
       x="county",
       y="incident_num")
```

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?
```{r}
sharks %>%
  select(incident_num, month) %>%
  count(month) %>%
  ggplot(aes(x=month, y=n))+
  geom_point(na.rm=T)+
  labs(title="Shark Incidents by Month",
       x="month",
       y="incident_num")
```

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?

```{r}
sharks %>%
  select(county, injury, incident_num) %>%
  group_by(county) %>%
  filter(injury=="fatal")
```
The county with the most fatal incidents is San Luis Obispo.
6. (2 points) In the data, `mode` refers to a type of activity. Which activity is associated with the highest number of incidents?
```{r}
sharks %>%
  select(incident_num, mode) %>%
  group_by(mode) %>%
  count(mode)

```
The most incidents is due to surfing and boarding.
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)
```{r}
sharks %>%
  select(mode, injury, incident_num) %>%
```

```
  group_by(mode) %>%
  ggplot(aes(x=mode, fill=injury))+
  geom_bar(position="dodge")+
  theme(axis.text.x=element_text(angle=45, hjust = 1))
```

8. (1 point) Which shark species is involved in the highest number of incidents?
```{r}
sharks
```

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.

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
```{r message=FALSE, warning=FALSE}
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?
```{r}
glimpse(white_sharks)
```

```{r}
str(white_sharks)
```

```{r}
any_na(white_sharks)
```
The missing values are represented by NAs.
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).
```{r}

```

12. (3 points) Make a plot that compares the range of total length by sex.
```{r}
white_sharks %>%
  select(total_length_cm, sex) %>%
  ggplot(aes(x=sex, fill=total_length_cm))+
  geom_bar(position="dodge")
```

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.