CrocWatch - Data Containing Crocodile Distributions in Queensland, Australia

By Tyler Lynch and Sam Tolar

Summary of the Data Set

- After clean up, we had 343 observations we could use to analyze the data
- The data consists of all of the crocodile sightings in Queensland, Australia, in 2020
- In case you did not know, Queensland is an Australian state, located in the North East of the continent
- Also in case you did not know, crocodiles are very large and dangerous reptilian animals

Why We Chose this Data Set

- Crocodiles have been my favorite animal since I was a kid
- This data set had so many variables as well as observations, we felt we could do a lot with it.

Our Relevant Variables

Longitude

Latitude

Species (Crocodylus johnstoni, Crocodylus porsus)

Length in Meters

Amount

Threat

Cleaning Up the Data

Our Process:

- We started with almost 2000 observations in the data set, but after clean up we were reduced to 343 observations
- Renaming relevant variables
- Focusing on relevant variables
- Converting data types from char to numeric/date
- Removing null values
- Converting yes/no columns to true/false
- Removing data errors

What Questions Did We Want to Ask?

What are the factors that correlate with a crocodile being labeled as threatening?

- Are crocodiles more threatening at certain times of the year?
- Are there certain areas where crocodiles are more threatening?
- Which species of crocodile were more likely to be found threatening?

Where were the crocodiles located?

- Where were the crocodiles located and how many were there?

What Questions Did We Want to Ask?

Can we predict the length of the crocodile based on the location, species, exposure, and if it is a threat or not?

Can we predict whether a crocodile is a threat based on location, species, exposure, length, month of sighting, and how many crocodiles were spotted at that time?

What range of values contain the true average of the length of crocodiles in Queensland, Australia?

How much larger are saltwater crocodiles than freshwater crocodiles?

```
> summary(data_cleaned)
      long            lat              date              species             meters       exposed          threat          amount       
 Min.   :138.8   Min.   :-28.109   Min.   :2020-01-01   Length:343         Min.   :0.300   Mode :logical   Mode :logical   Min.   : 1.000  
 1st Qu.:145.7   1st Qu.:-19.257   1st Qu.:2020-03-08   Class :character   1st Qu.:2.000   FALSE:75        FALSE:150       1st Qu.: 1.000  
 Median :146.0   Median :-17.497   Median :2020-06-12   Mode  :character   Median :3.000   TRUE :268       TRUE :193       Median : 1.000  
 Mean   :146.4   Mean   :-18.001   Mean   :2020-06-19                      Mean   :2.751                                   Mean   : 1.134  
 3rd Qu.:146.8   3rd Qu.:-16.776   3rd Qu.:2020-09-12                      3rd Qu.:3.500                                   3rd Qu.: 1.000  
 Max.   :153.4   Max.   : -9.756   Max.   :2020-12-12                      Max.   :6.000                                   Max.   :12.000  
>
```

Crocodylus johnstoni	Crocodylus porsus
13	330

Frequency table of species

Species	Exposed	
	FALSE	TRUE
Crocodylus johnstoni	3	10
Crocodylus porsus	72	258

Two Way Contingency table showing the species and whether it is fully exposed

Species	Threat	
	FALSE	TRUE
Crocodylus johnstoni	9	4
Crocodylus porsus	141	189

Two Way Contingency table showing the species and whether it was considered a threat

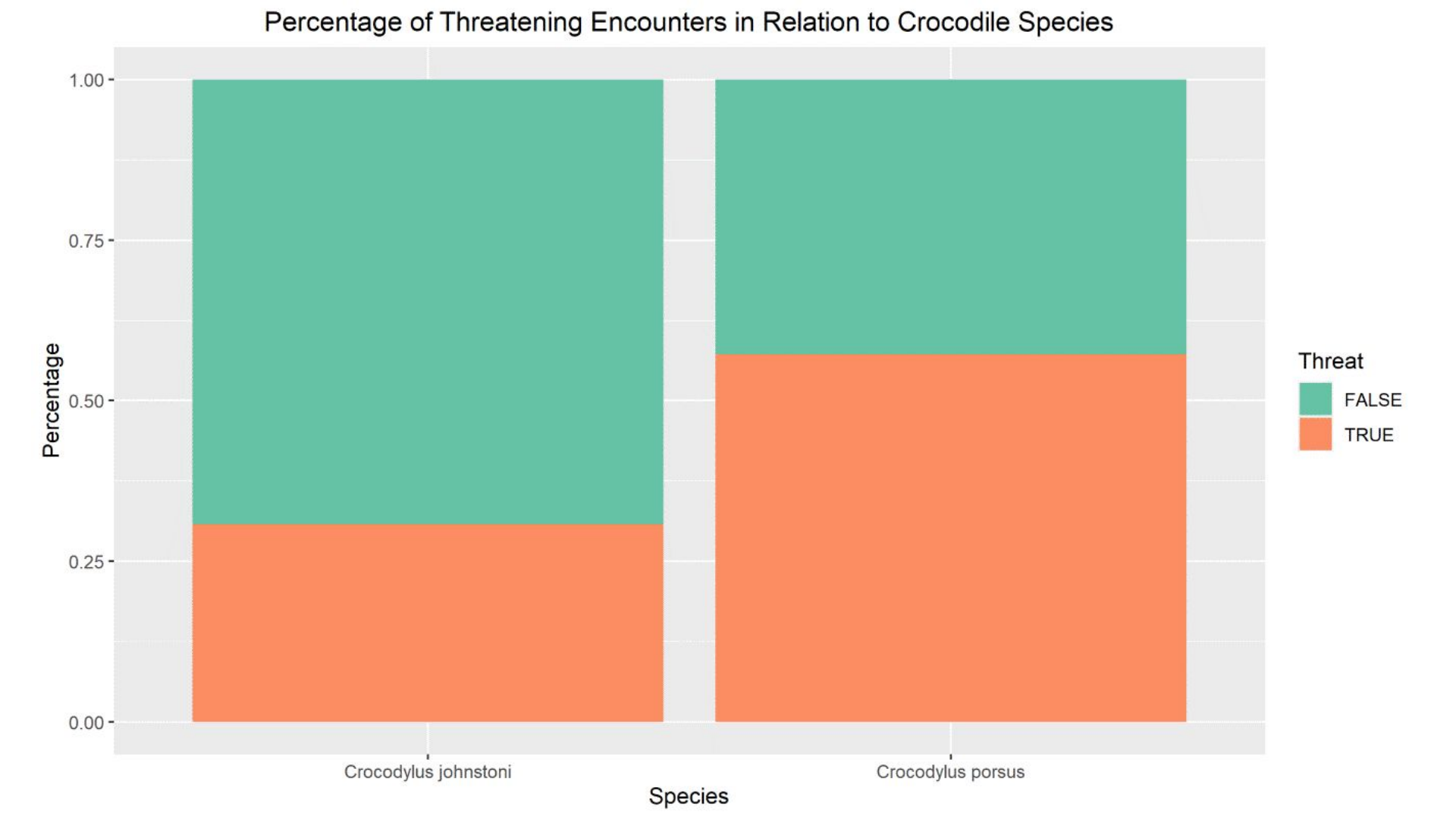
Exposed	Threat	
	FALSE	TRUE
FALSE	28	47
TRUE	122	146

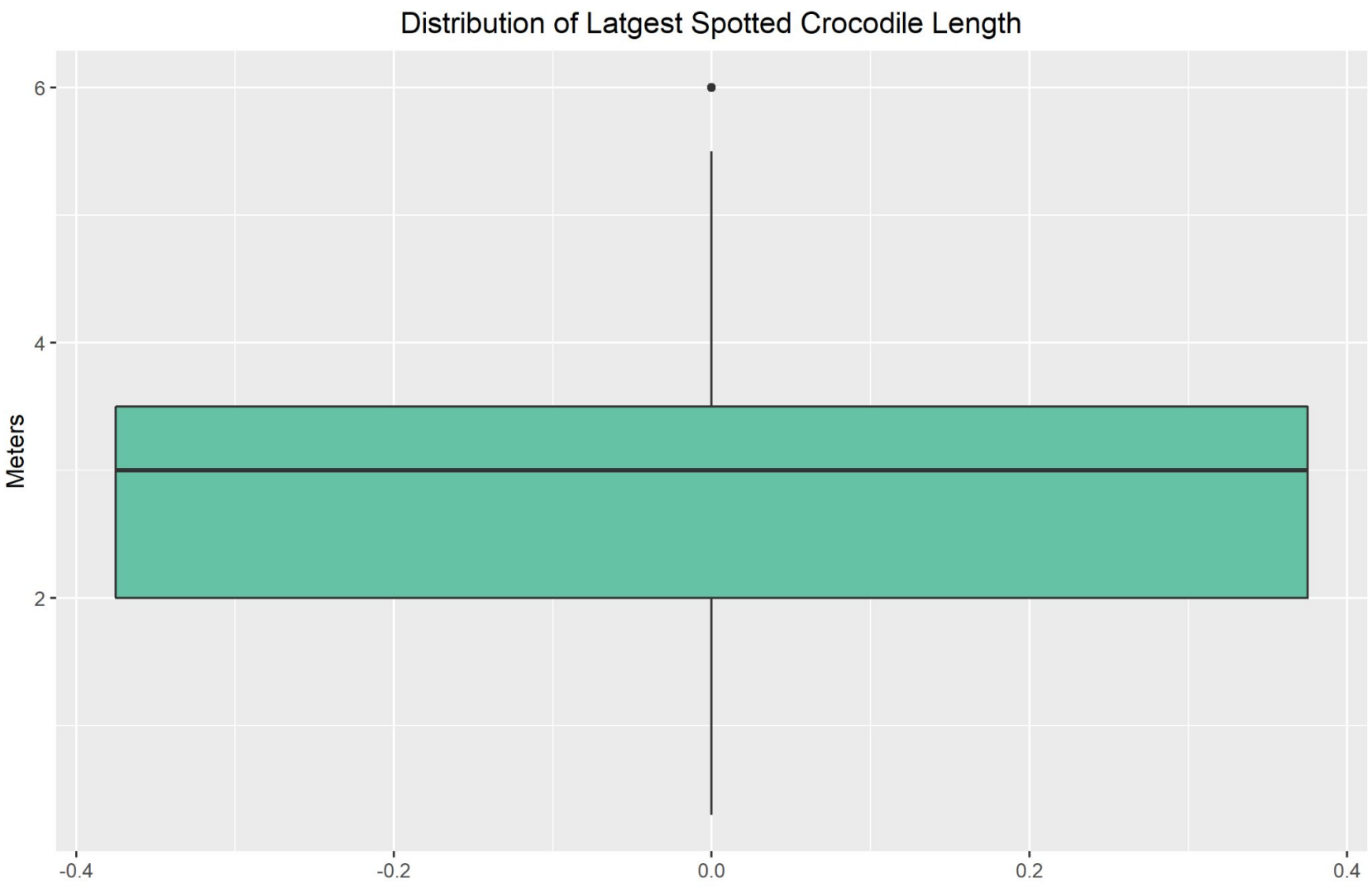
Two Way Contingency table of whether the crocodile was fully exposed and if it was considered a threat

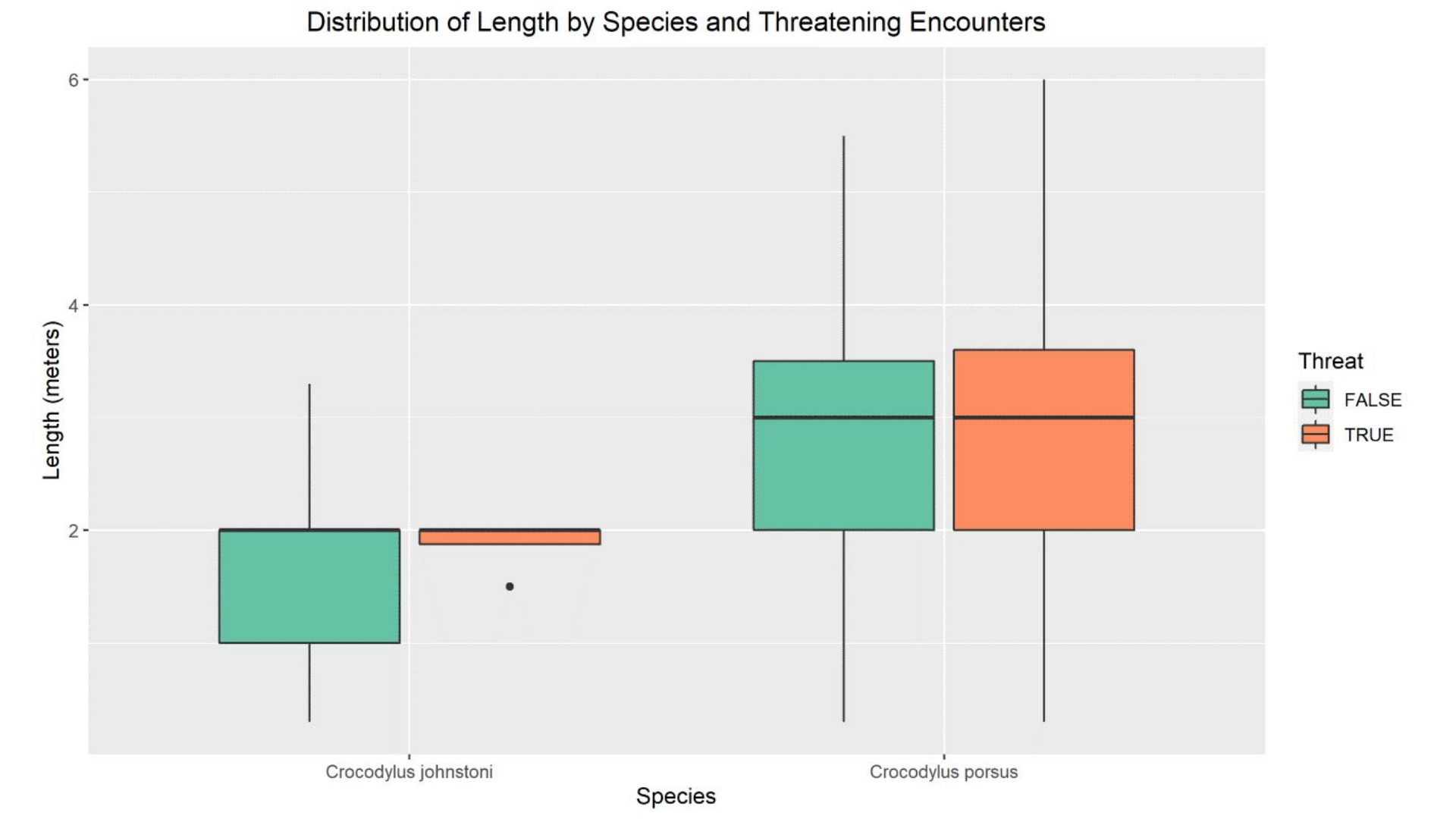
Data Visualization

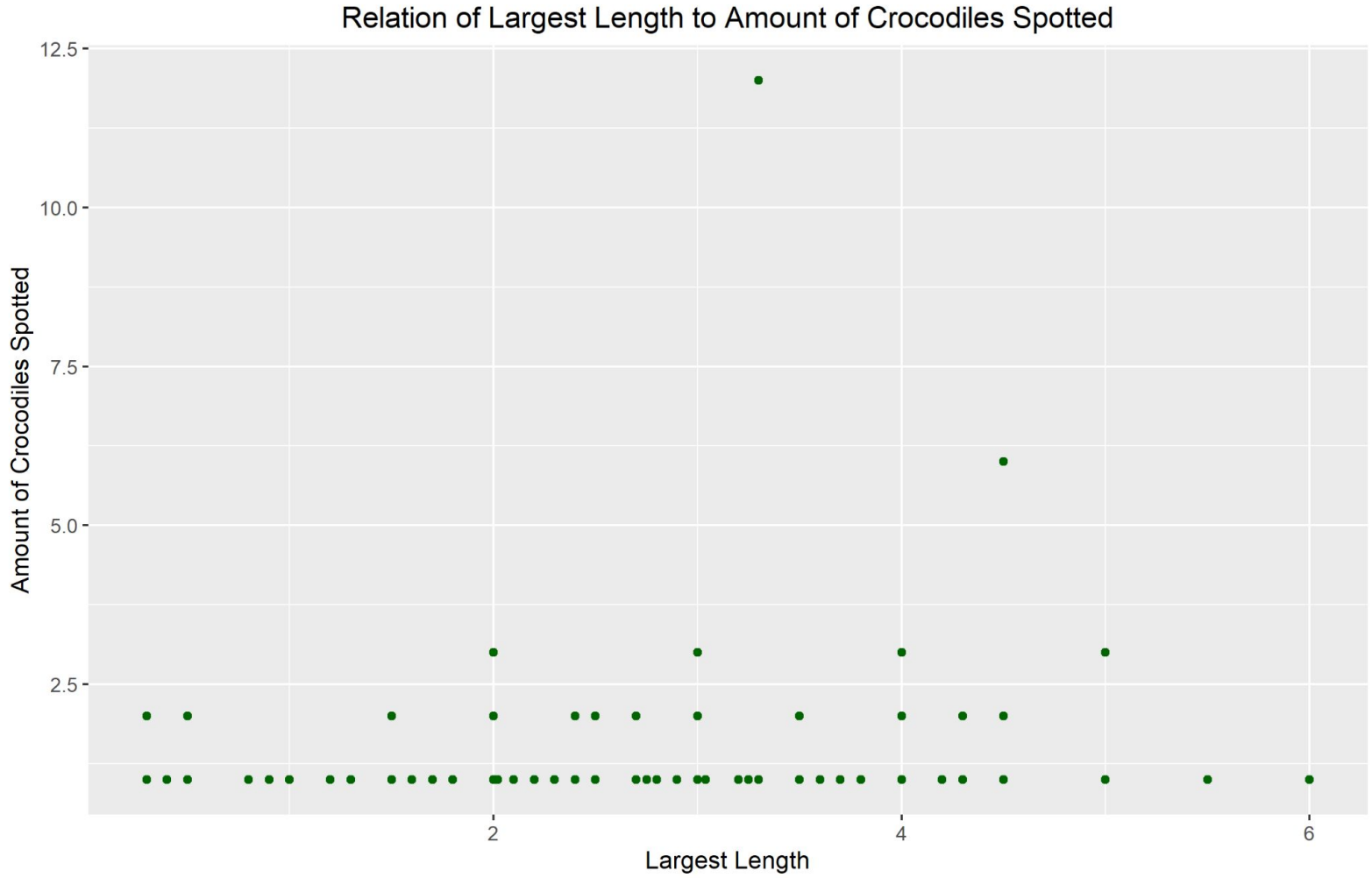
We made..

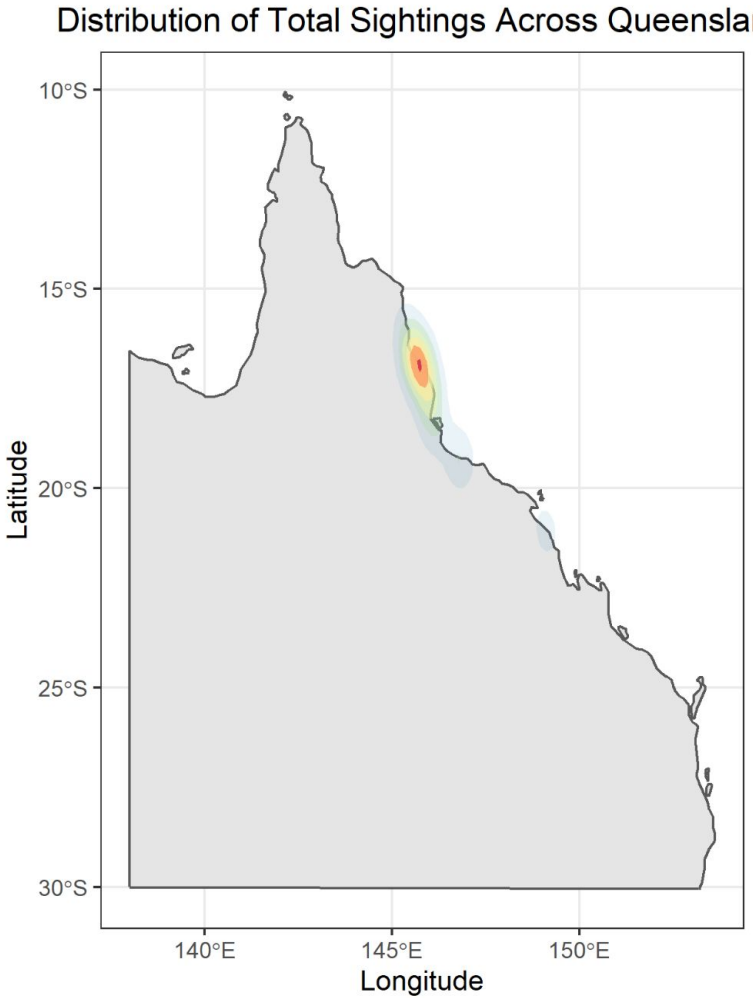
- A bar graph of categorical variables (species related to threat)
- A boxplot showing the distribution of the length of the crocodiles
- A side by side boxplot showing the length of the crocodile by species
- A scatterplot showing the relation of the largest length of a crocodile in a group to how many were spotted
- A map showing the location of threatening and non-threatening species by latitude and longitude
- A heat map showing the distribution of total sightings across Queensland
- A time series of total number of threatening crocodiles per month
- A map of sightings across Queensland showing the number of crocodiles per sighting

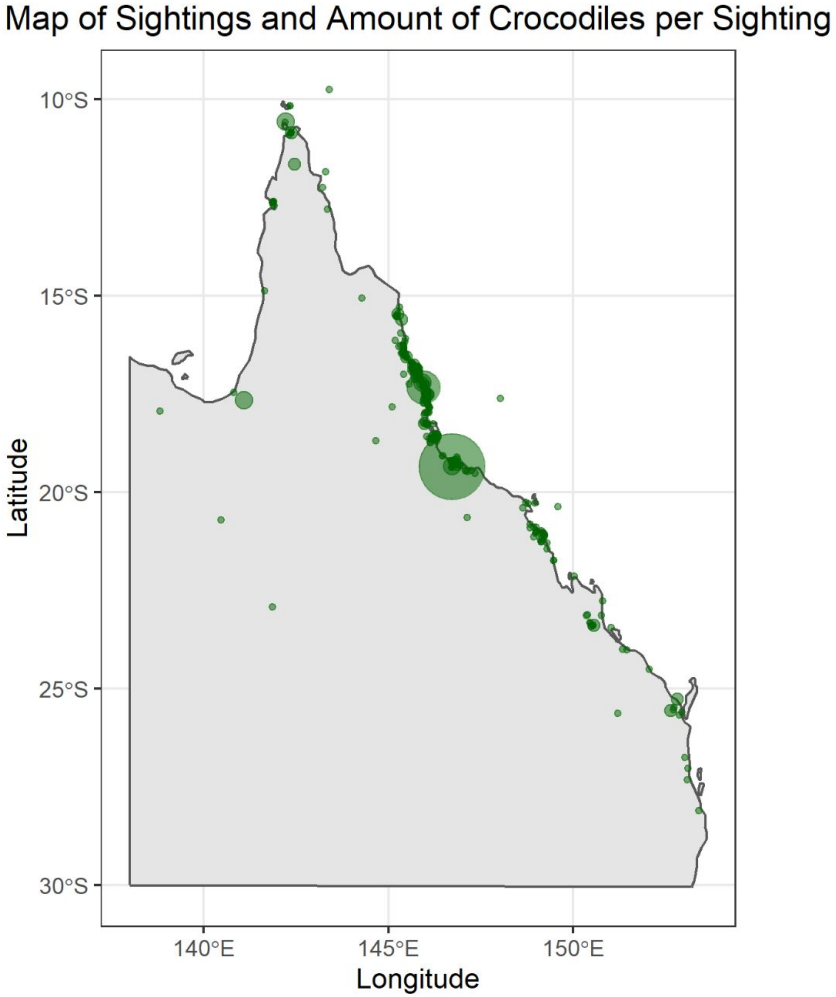

Percentage of Threatening Encounters in Relation to Crocodile Species
1.00
0.75
0.50
0.25
0.00
Percentage
Crocodylus johnstoni
Crocodylus porsus
Species
Threat
FALSE
TRUE

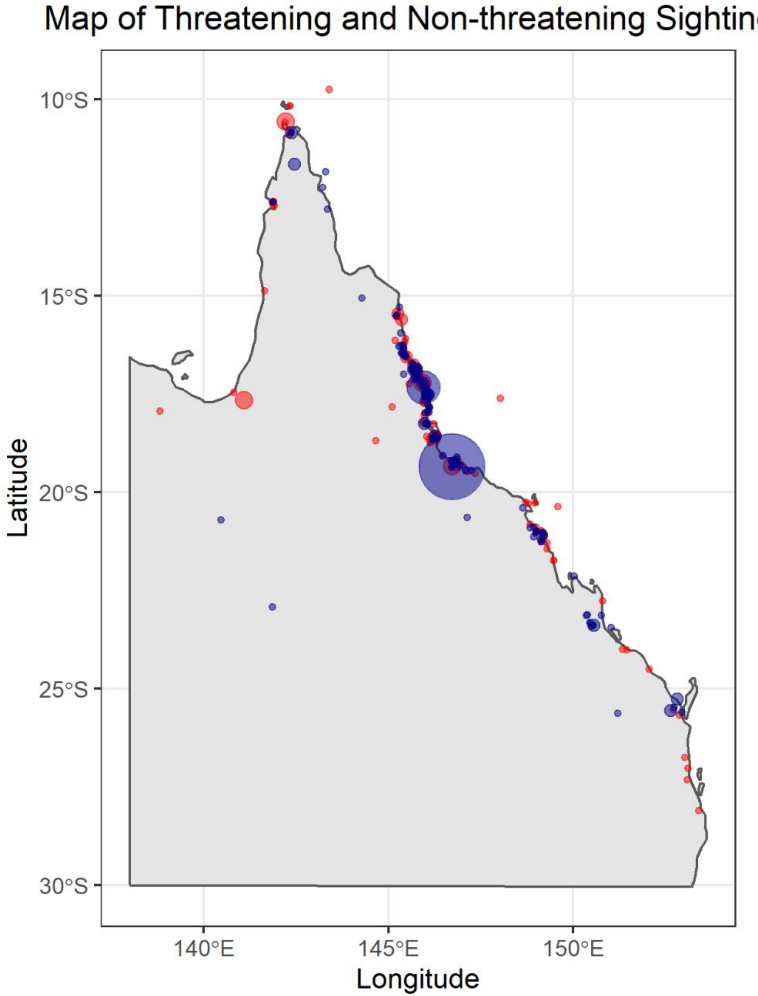

Distribution of Latgest Spotted Crocodile Length
Meters
6
4
2
-0.4
-0.2
0.0
0.2
0.4

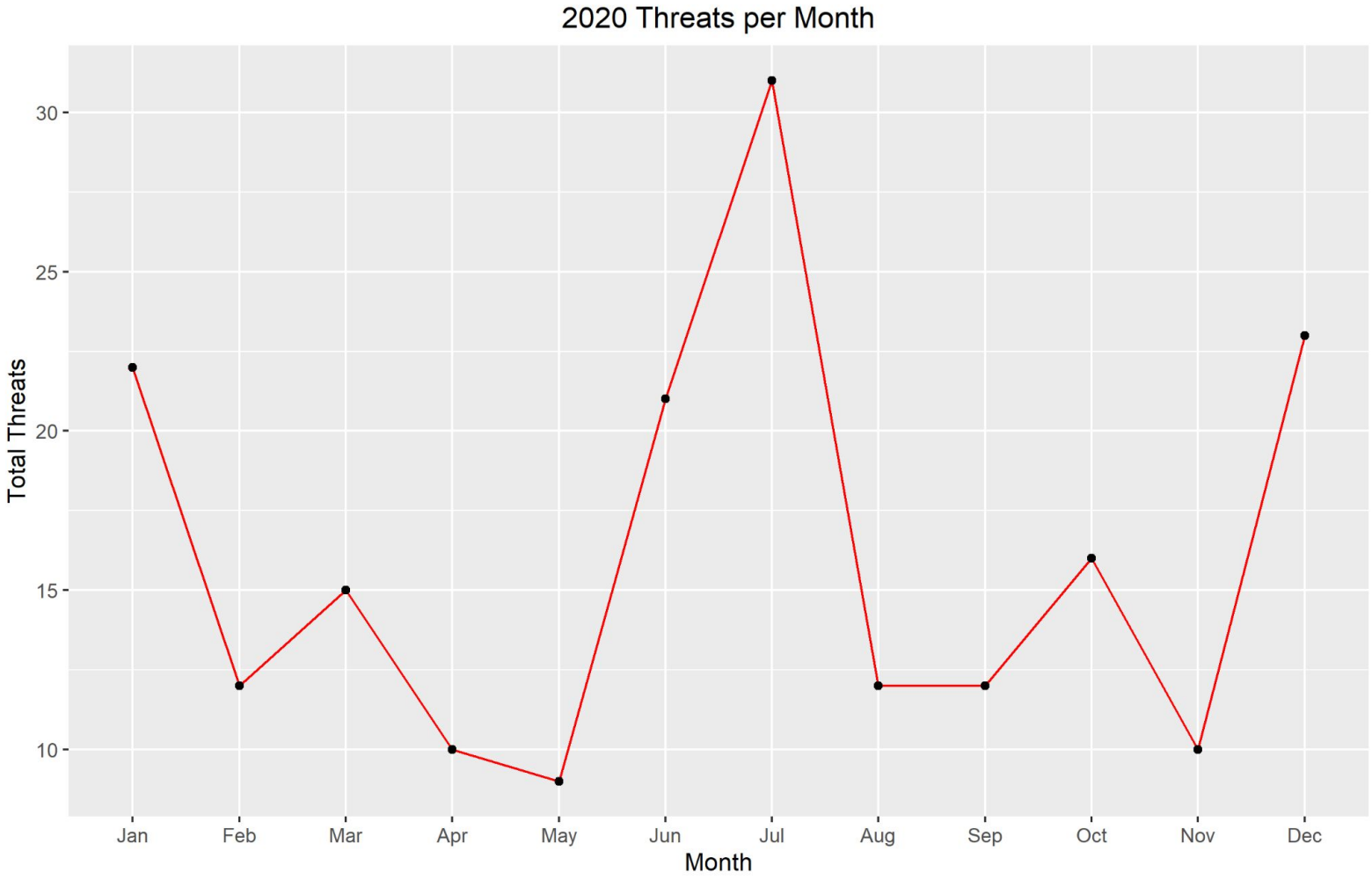

Distribution of Length by Species and Threatening Encounters
Length (meters)
6
4
2
Crocodylus johnstoni
Crocodylus porsus
Species
Threat
FALSE
TRUE


Relation of Largest Length to Amount of Crocodiles Spotted
Amount of Crocodiles Spotted
12.5
10.0
7.5
5.0
2.5
2
4
6
Largest Length


Distribution of Total Sightings Across Queensland
10°S
15°S
20°S
25°S
30°S
Latitude
140°E
145°E
150°E
Longitude


Map of Sightings and Amount of Crocodiles per Sighting
10°S
15°S
20°S
25°S
30°S
Latitude
140°E
145°E
150°E
Longitude


Map of Threatening and Non-threatening Sightings
10°S
15°S
20°S
25°S
30°S
Latitude
140°E
145°E
150°E
Longitude


2020 Threats per Month
Total Threats
30
25
20
15
10
Jan
Feb
Mar
Apr
May
Jun
Jul
Aug
Sep
Oct
Nov
Dec
Month

One and Two Sample Inference

What range of values contain the true average of the length of crocodiles in Queensland, Australia?

How much larger are saltwater crocodiles than freshwater crocodiles?

One-Sample Confidence Interval of the Length of Crocodiles

```
> # One-sample confidence interval of length of crocodiles
> t.test(data_cleaned$meters, conf.level = 0.95)

        One Sample t-test

data:  data_cleaned$meters
t = 46.426, df = 342, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 2.634357 2.867450
sample estimates:
mean of x
 2.750904
```

Two-Sample Confidence Interval of the Difference in Lengths Between Freshwater and Saltwater Crocodiles

```
> # Two-sample confidence interval of the
> t.test(meters~species, conf.level=0.95, data=data_cleaned)

        Welch Two Sample t-test

data:  meters by species
t = -4.6931, df = 13.852, p-value = 0.0003555
alternative hypothesis: true difference in means between group Crocodylus johnstoni and group Crocodylus porsus is not equal to
 0
95 percent confidence interval:
 -1.5570354 -0.5795986
sample estimates:
mean in group Crocodylus johnstoni    mean in group Crocodylus porsus
                          1.723077                           2.791394
```

Regression

Just a refresher-

Can we predict the length of the crocodile based on the location, species, exposure, and if it is a threat or not?

Can we predict whether a crocodile is a threat based on location, species, exposure, length, month of sighting, and how many crocodiles were spotted at that time?

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             13.28454    8.38250   1.585 0.113951
long                    -0.08497    0.06329  -1.343 0.180297
lat                     -0.04770    0.04869  -0.980 0.327935
speciesCrocodylus porsus  1.16267    0.33001   3.523 0.000485 ***
exposedTRUE             -0.20318    0.14257  -1.425 0.155035
threatTRUE               0.14883    0.11888   1.252 0.211470
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.077 on 337 degrees of freedom
Multiple R-squared:  0.05129,   Adjusted R-squared:  0.03722
F-statistic: 3.644 on 5 and 337 DF,  p-value: 0.003177
```

Initial Linear Regression Model

AIC Linear Regression Model

```
Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)               1.6697     0.3012   5.544 5.94e-08 ***
speciesCrocodylus porsus  1.0223     0.3064   3.337 0.000942 ***
threatTRUE                0.1735     0.1179   1.471 0.142140
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.078 on 340 degrees of freedom
Multiple R-squared:  0.04077,   Adjusted R-squared:  0.03512
F-statistic: 7.225 on 2 and 340 DF,  p-value: 0.0008456
```

Linear AIC Model Prediction

```
> # Prediction of the length of a threatening Crocodylus porus that is a threat:
> predict(lm_AICmodel, data.frame(species="Crocodylus porsus", threat=T), interval="predict")
      fit       lwr      upr
1 2.86554 0.7397473 4.991333
```

Initial Logistic Regression Model

```
Coefficients:

(Intercept)
long
lat
speciesCrocodylus porsus
meters
exposedTRUE
amount
month
```

AIC Logistic Regression Model

(Intercept)	long	speciesCrocodylus porsus	exposedTRUE
4.019257e+06	8.979893e-01	3.254999e+00	6.482245e-01

Logistic AIC Model Prediction

```
> # Prediction of whether a crocodile sighting of a Crocodylus porsus at longitude 145
> # that is fully exposed is a threat or not:
> predict(glm_AICmodel, data.frame(species="Crocodylus porsus", long=145, exposed=T), type="response")
        1
0.5870282
```

Citations

https://www.data.qld.gov.au/dataset/crocodile-sightings-in-queensland/resource/6b0e71dd-4148-4934-b919-d50935d14417

https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.foxnews.com%2Fgreat-outdoors%2Faustralian-fisherman-crocodile-home&psig=AOvVaw39sM22kZLM5piABAW3Dpxf&ust=1638859687902000&source=images&cd=vfe&ved=0CAwQjhxqFwoTCOjv_qLKzvQCFQAAAAAdAAAAABAD

https://www.google.com/url?sa=i&url=https%3A%2F%2Fen.wikipedia.org%2Fwiki%2FQueensland&psig=AOvVaw2SwecmFzqdmqIjJ9RBoE48&ust=1638859725647000&source=images&cd=vfe&ved=0CAsQjRxqFwoTCJiV6rHKzvQCFQAAAAAdAAAAABAD

https://axelhodler.medium.com/creating-a-heat-map-from-coordinates-using-r-780db4901075/

https://www.r-bloggers.com/2019/04/zooming-in-on-maps-with-sf-and-ggplot2/

GitHub with all of our code and graphs:

https://github.com/tyler-lynch/Crocodile-Analysis