# dolphin_strandings

Emma Beyer & Ayoung Kim

2024-02-27

```
#loading packages
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.3     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(rmarkdown)
#setting WD
#setwd("/home/guest/Stat_Modeling_Lab/ENV710Ayoung_Emma")

##original_data <- read.csv("dolphins.csv")

#read in data for problems 1-2
strandings <- read.csv("./dolphins_cleaned.csv")
#attach data
attach(strandings)
```

```
#remove NAs
cleaned_strandings <- na.omit(strandings)

#remove unneeded covariates
cleaned_strandings <- subset(cleaned_strandings, select = -c(Shot, Fishery.Interaction, Boat.Collision))

#removing blanks in Age.Class
cleaned_strandings <- cleaned_strandings[!grepl("^\\s*$", cleaned_strandings$Age.Class), ]
```

```r
#removing unknowns in Age.Class
cleaned_strandings <- cleaned_strandings[!cleaned_strandings$Age.Class %in% c("UNKNOWN"), ]
#removing unknowns in Sex
cleaned_strandings <- cleaned_strandings[!cleaned_strandings$Sex %in% c("UNKNOWN"), ]

#create binary covariate for wind turbine presence in state
cleaned_strandings <- cleaned_strandings %>%
  mutate(turbine_presence = if_else(State %in% c("VA", "NY", "RI", "MA"), 1, 0))

write.csv(cleaned_strandings, "cleaned_strandings", row.names = FALSE)
#strandings_test <- read.csv("cleaned_strandings")


#creating subset using collision observations
subset <- cleaned_strandings[!grepl("^\\s*$", cleaned_strandings$Boat.Collision), ]

#create new csv file for future editing
subset <- subset[!subset$Boat.Collision %in% c("C"), ]

#count of strandings in wind farm states = 6
sum(subset$turbine_presence)
```

```
## [1] 0
```

```r
#count of strandings in turbine states = 11
sum(cleaned_strandings$turbine_presence)
```

```
## [1] 11
```

## Data Cleaning

```r
#remove unneeded covariates
cleaned_strandings_test <- subset(strandings, select = -c(Shot, Fishery.Interaction, Boat.Collision, We

#remove NAs
cleaned_strandings_test <- na.omit(cleaned_strandings_test)

#removing blanks in Age.Class
cleaned_strandings_test <- cleaned_strandings_test[!grepl("^\\s*$", cleaned_strandings_test$Age.Class),
#removing unknowns in Age.Class
cleaned_strandings_test <- cleaned_strandings_test[!cleaned_strandings_test$Age.Class %in% c("UNKNOWN")
#removing unknowns in Sex
cleaned_strandings_test <- cleaned_strandings_test[!cleaned_strandings_test$Sex %in% c("UNKNOWN"), ]

#create binary covariate for wind turbine presence in state
cleaned_strandings_test <- cleaned_strandings_test %>%
  mutate(turbine_presence = if_else(State %in% c("VA", "NY", "RI", "MA"), 1, 0))

#count of strandings in wind farm states = 80
sum(cleaned_strandings_test$turbine_presence)
```

```
## [1] 80
```

```
#create offshore wind subset
turbine_data <- cleaned_strandings_test[cleaned_strandings_test$State %in% c("VA", "NY", "RI", "MA"), ]

#covarients used: states, turbine presence, sex, length, age class

#States info
table(cleaned_strandings_test$State)
```

```
##
##  AL  DE  FL  GA  LA  MA  MD  ME  MS  NC  NJ  NY  RI  SC  TX  VA
##  85  25 295  39  92   9  20   1 157 182  35  17   1 102 306  53
```

```
# 15 states included
# Most strandings were in Florida (295) and Texas (306)
# Least were in Maine (1) and Rhode Island (1)
# States with offshore wind: Virginia, New York, Rhode Island, and Massachusetts

#count of strandings in wind farm states = 80
sum(cleaned_strandings_test$turbine_presence)
```

```
## [1] 80
```

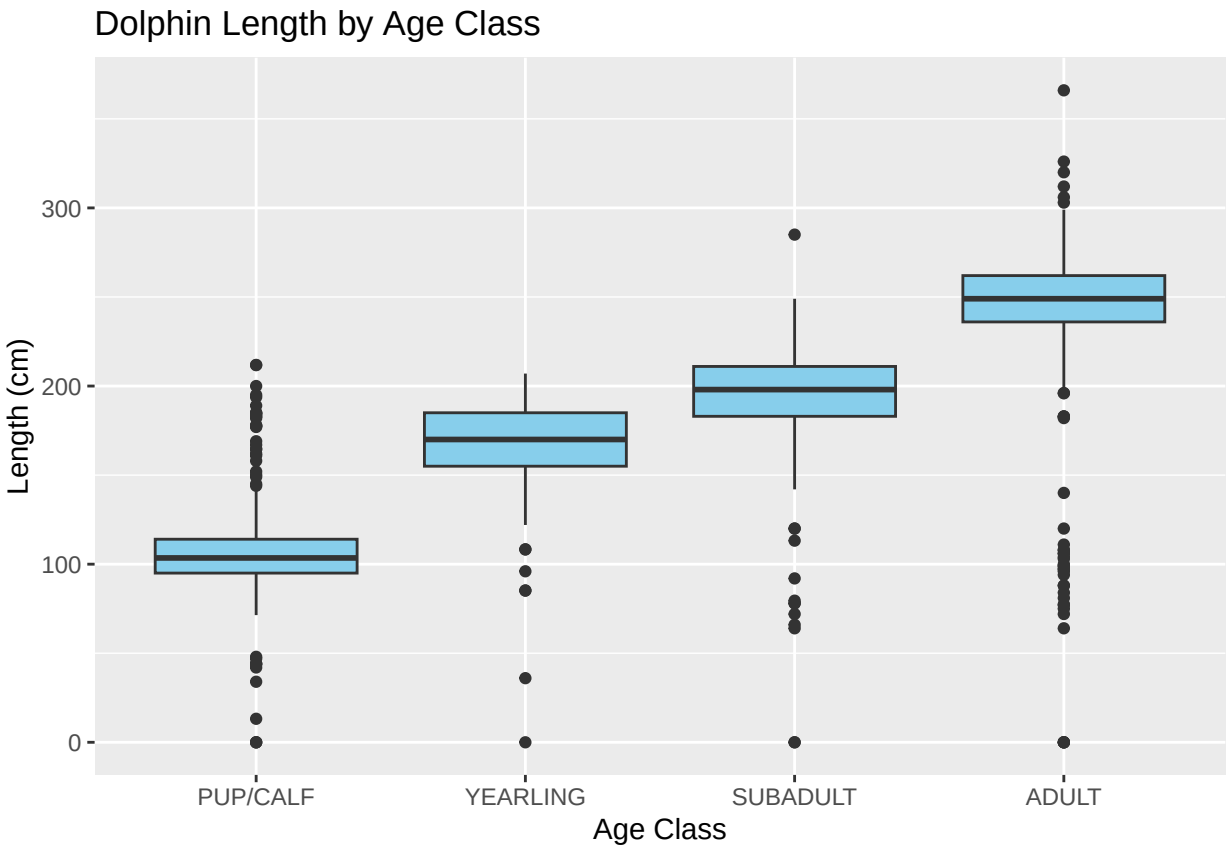
```
#Length info
# average overall length 194.2091
mean(cleaned_strandings_test$Length)
```

```
## [1] 194.2091
```

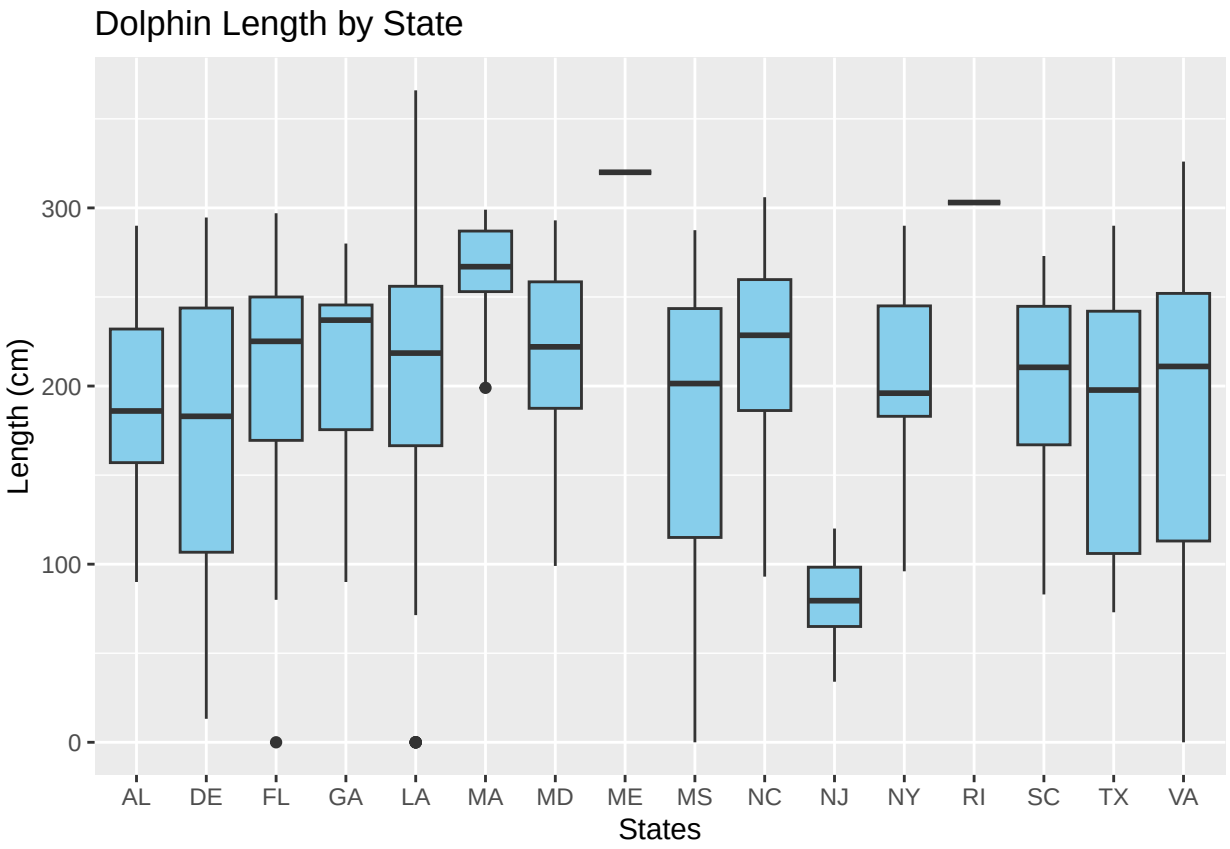
```
# average offshore wind length = 198.175
mean(turbine_data$Length)
```

```
## [1] 198.175
```

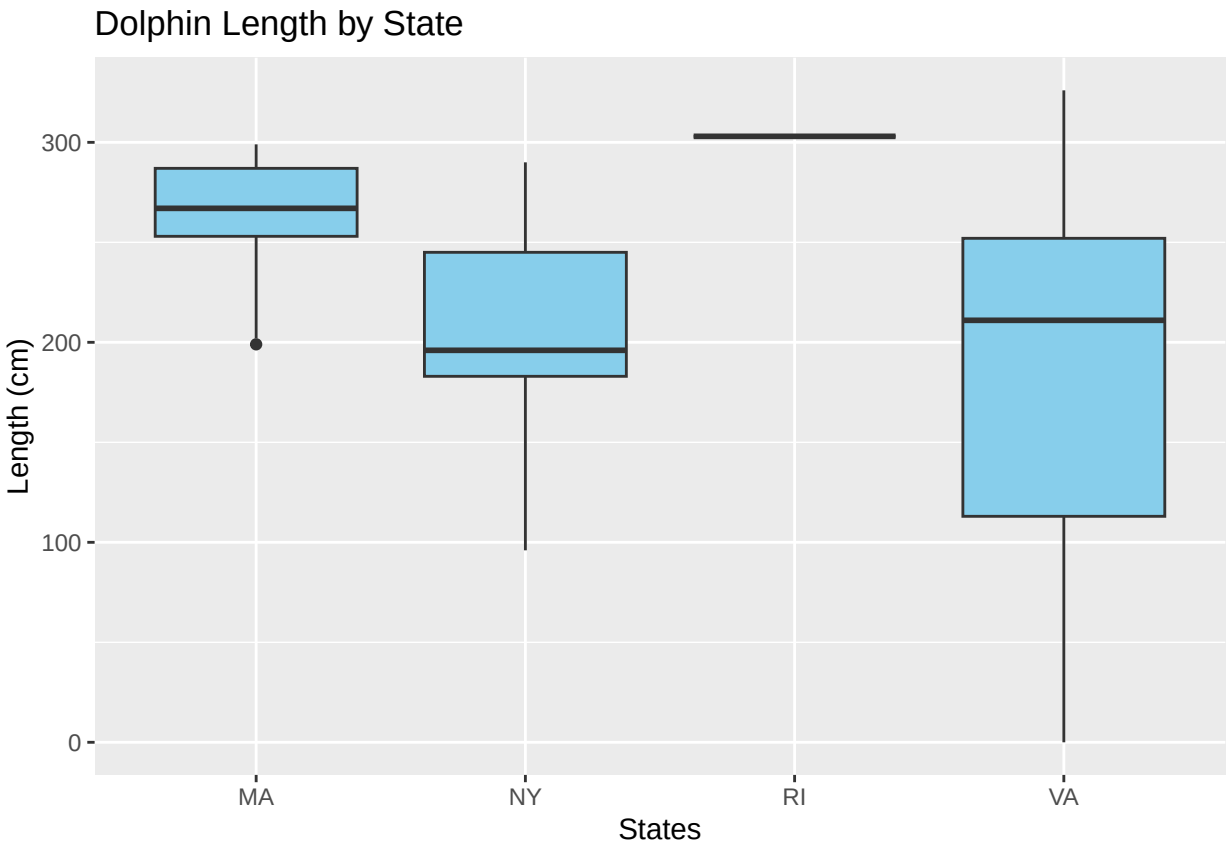
```
#reordering age classes so they are from youngest to oldest
age_class_order <- c("PUP/CALF", "YEARLING", "SUBADULT", "ADULT")

#summary graph of Age Class by Length
length_ageclass <- ggplot(cleaned_strandings_test, aes(x = factor(Age.Class, levels = age_class_order),
  geom_boxplot(fill="skyblue") +
  labs(x = "Age Class", y = "Length (cm)") +
  ggtitle("Dolphin Length by Age Class")
length_ageclass
```



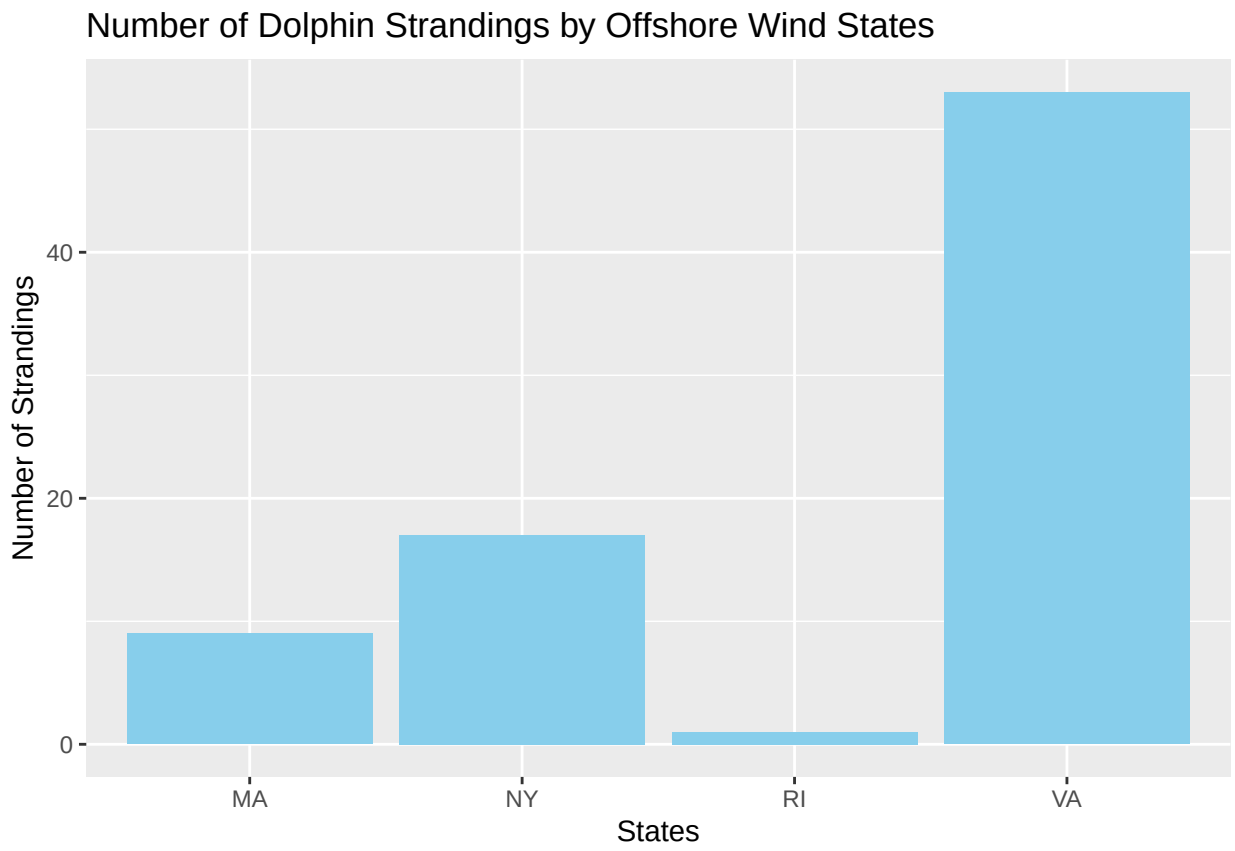

```
#plot of Lengths in each state
state_length <- ggplot(cleaned_strandings_test, aes(x = State, y = Length)) +
  geom_boxplot(fill="skyblue") +
  labs(x = "States", y = "Length (cm)") +
  ggtitle("Dolphin Length by State")
state_length
```



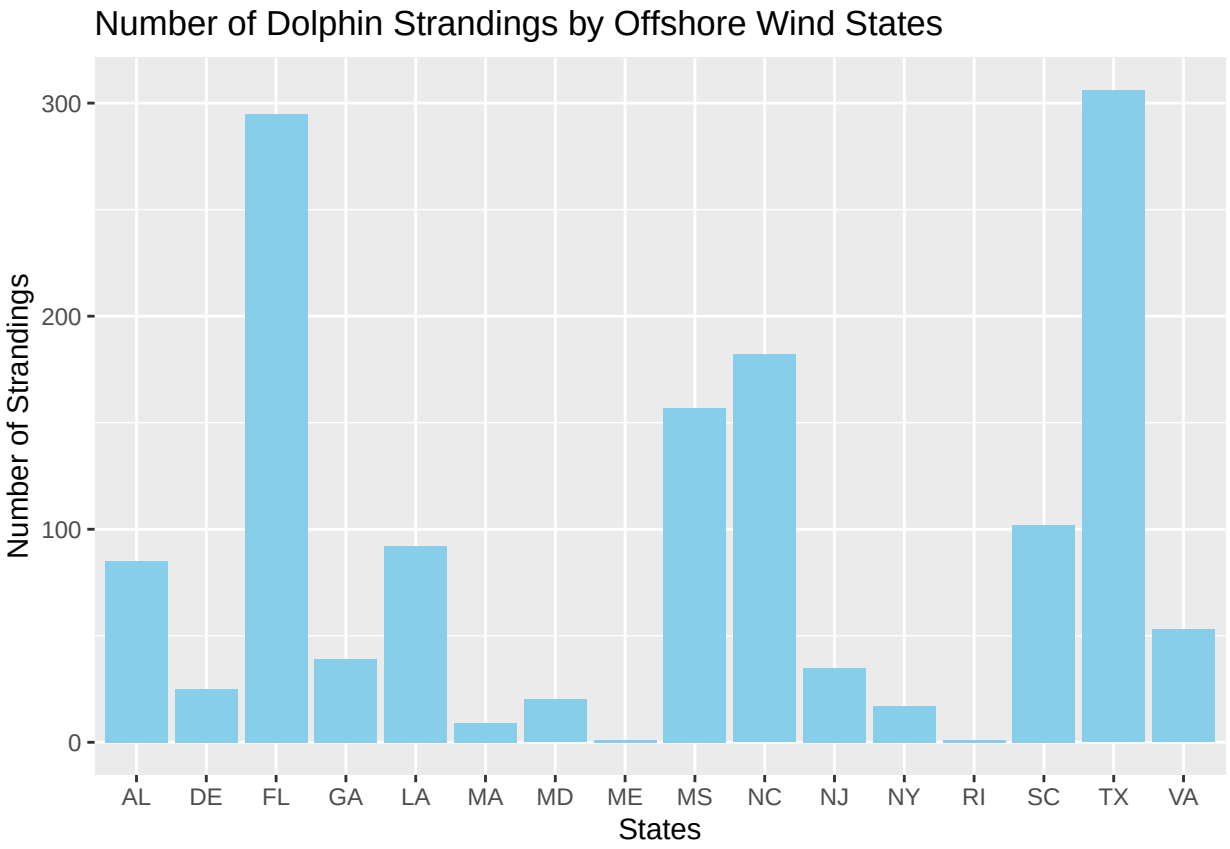

```
#plot of Lengths in each offshore wind state
turbinestate_length <- ggplot(turbine_data, aes(x = State, y = Length)) +
  geom_boxplot(fill="skyblue") +
  labs(x = "States", y = "Length (cm)") +
  ggtitle("Dolphin Length by State")
turbinestate_length
```



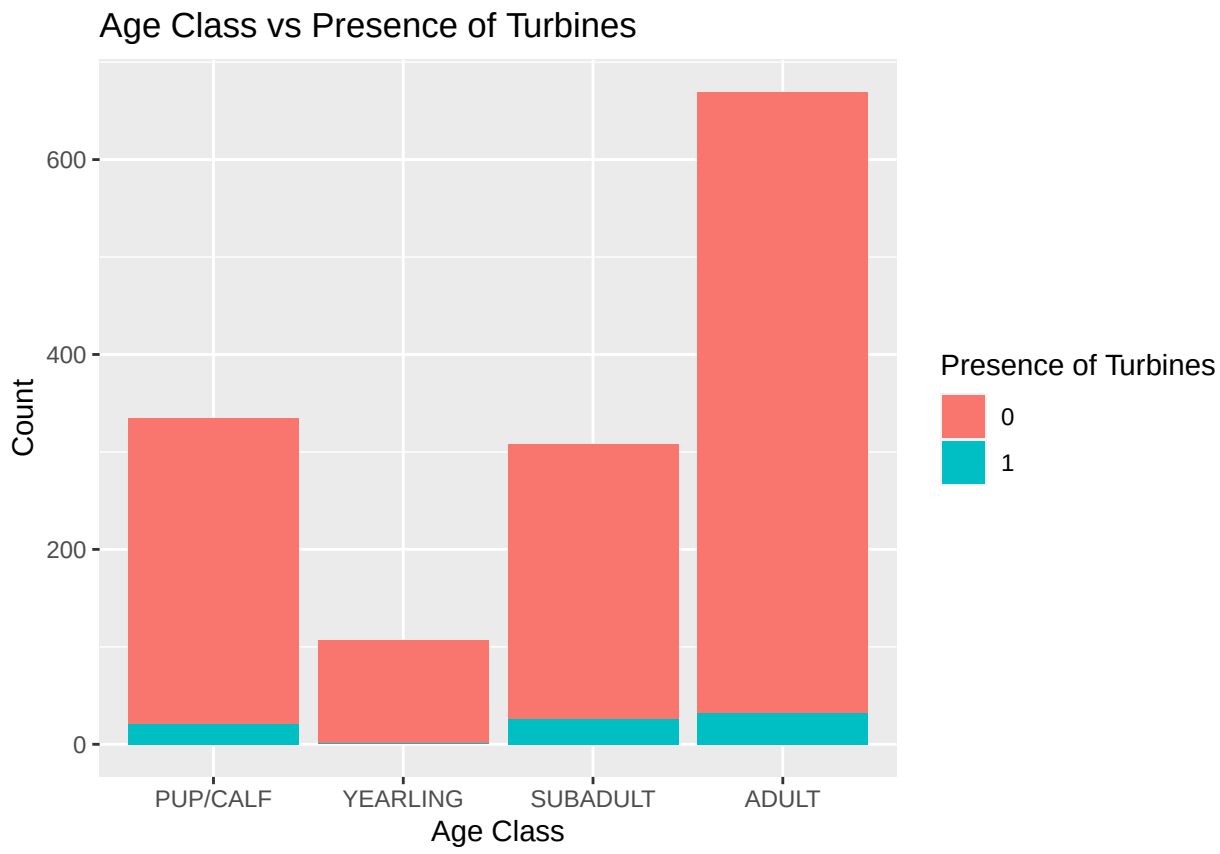

```
#plot of number of strandings for each offshore wind state
standings_wf_state <- ggplot(turbine_data, aes(x = State)) +
  geom_bar(fill = "skyblue") +
  labs(x = "States", y = "Number of Strandings") +
  ggtitle("Number of Dolphin Strandings by Offshore Wind States")
standings_wf_state
```



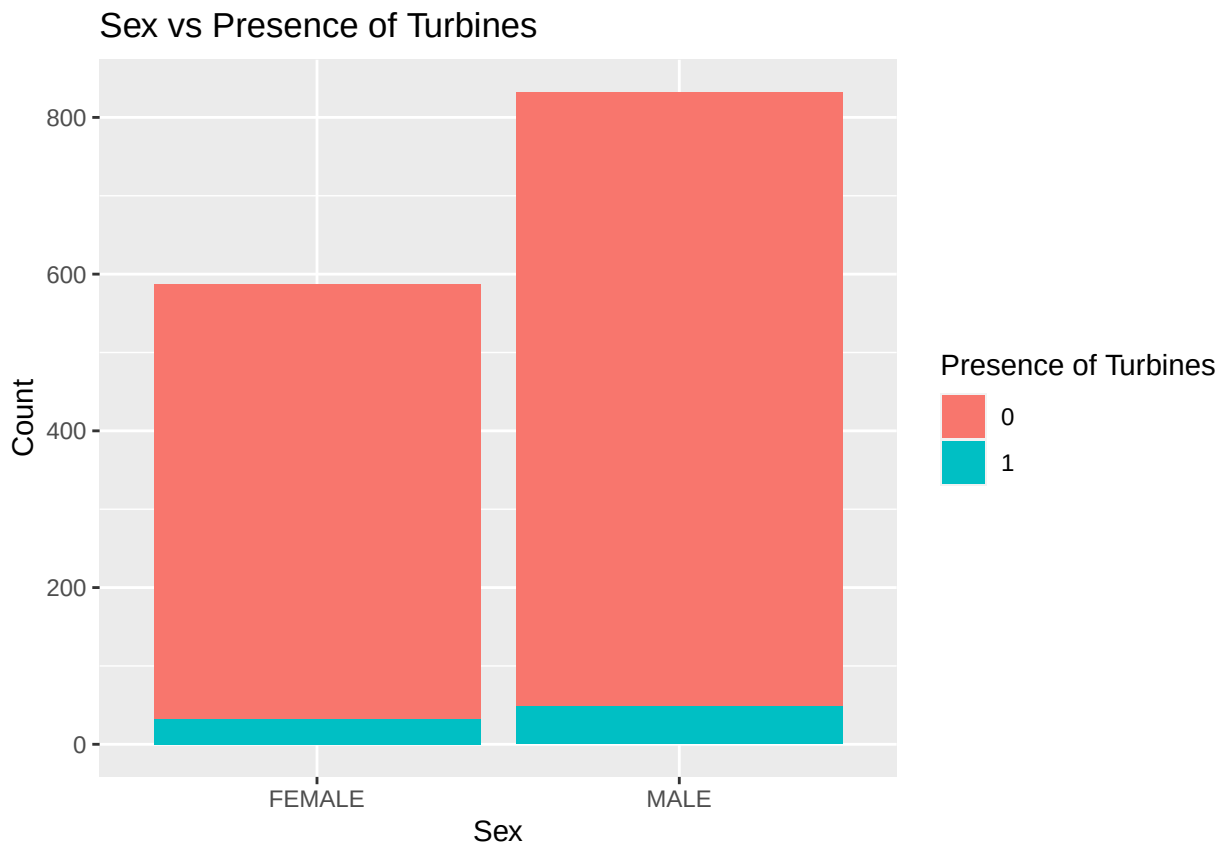

```
#plot of number of strandings for each offshore wind state
standings_state <- ggplot(cleaned_strandings_test, aes(x = State)) +
  geom_bar(fill = "skyblue") +
  labs(x = "States", y = "Number of Strandings") +
  ggtitle("Number of Dolphin Strandings by Offshore Wind States")
standings_state
```




```
#plot of the number of strandings in each age class
strandings_ageclass <- ggplot(cleaned_strandings_test, aes(x = factor(Age.Class, levels = age_class_ord
  geom_bar(position = "stack") +
  labs(x = "Age Class", y = "Count", fill = "Presence of Turbines") +
  ggtitle("Age Class vs Presence of Turbines")
strandings_ageclass
```




```
#plot of the number of strandings in each sex
strandings_sex <- ggplot(cleaned_strandings_test, aes(x = factor(Sex), fill = factor(turbine_presence)))
  geom_bar(position = "stack") +
  labs(x = "Sex", y = "Count", fill = "Presence of Turbines") +
  ggtitle("Sex vs Presence of Turbines")
strandings_sex
```




```
#Fitting a regression model (Turbine Presence)
fit_1 <- lm(turbine_presence~1, data = cleaned_strandings)
#Summary of the regression model fit_1
summary(fit_1)
```

```
##
## Call:
## lm(formula = turbine_presence ~ 1, data = cleaned_strandings)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.02124 -0.02124 -0.02124 -0.02124  0.97876
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.021236   0.006341   3.349  0.00087 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1443 on 517 degrees of freedom
```

```
#Checking the structure of Cleaned_strandings
str(cleaned_strandings)
```

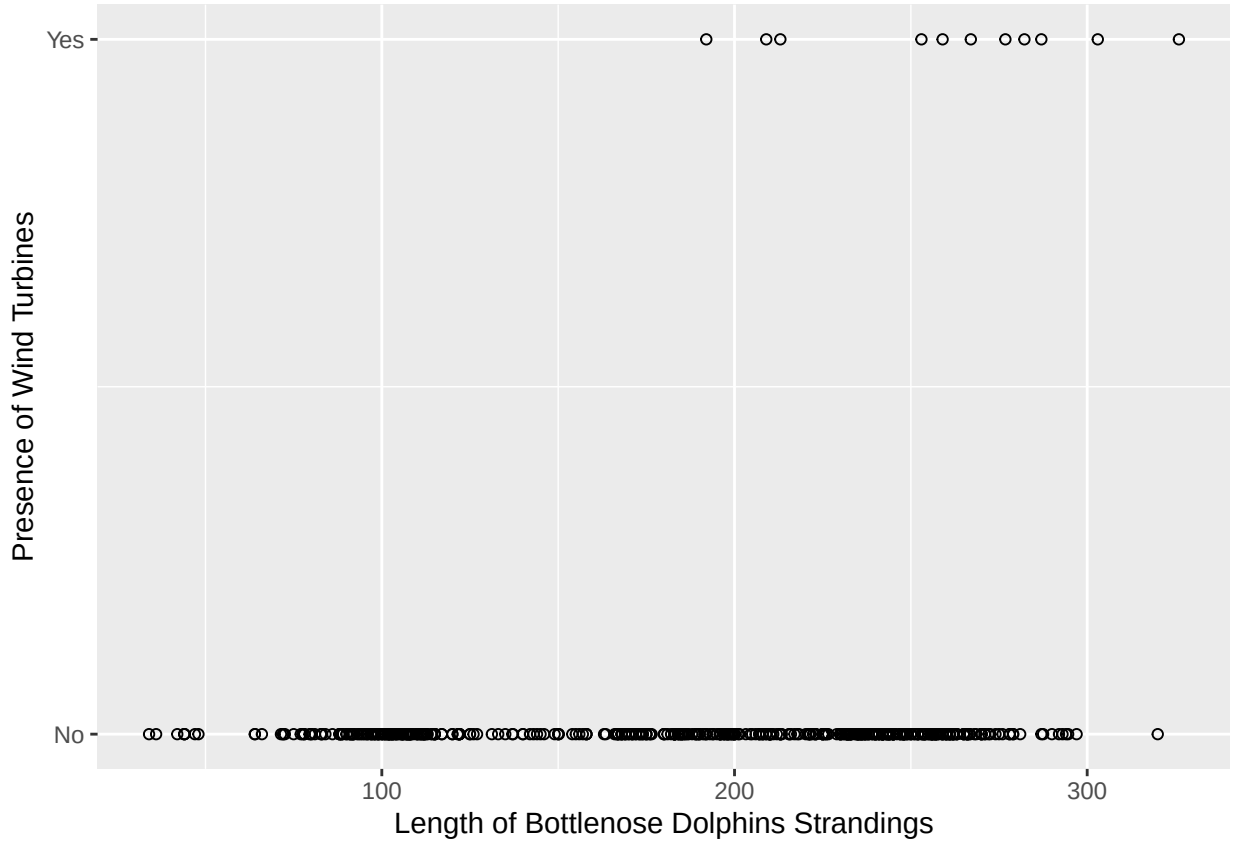
```
## 'data.frame':    518 obs. of  7 variables:
```

```
##  $ State              : chr  "AL" "AL" "AL" "AL" ...
##  $ Year.of.Observation: int  2017 2017 2019 2017 2017 2019 2018 2018 2019 2017 ...
##  $ Sex                : chr  "MALE" "MALE" "MALE" "MALE" ...
##  $ Age.Class          : chr  "SUBADULT" "SUBADULT" "ADULT" "ADULT" ...
##  $ Length             : num  230 225 249 263 279 ...
##  $ Weight             : num  161 122 153 221 229 78 17.2 10 12 80.9 ...
##  $ turbine_presence   : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
#Fitting a regression model (finding the odds of different sex in offshore wind states)
fit_2 <- glm(turbine_presence~Sex,family='binomial', data = cleaned_strandings)
#Summary of the regression model fit_2
summary(fit_2)
```

```
##
## Call:
## glm(formula = turbine_presence ~ Sex, family = "binomial", data = cleaned_strandings)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -20.57    1164.06  -0.018    0.986
## SexMALE        17.35    1164.06   0.015    0.988
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106.51  on 517  degrees of freedom
## Residual deviance:  93.25  on 516  degrees of freedom
## AIC: 97.25
##
## Number of Fisher Scoring iterations: 19
```

```
#Plot for fit_1
fit_2_plot <- ggplot(cleaned_strandings, aes(x =Length, y =turbine_presence)) +
geom_point(shape = 1) +
xlab("Length of Bottlenose Dolphins Strandings") +
ylab("Presence of Wind Turbines") +
scale_y_continuous(breaks = c(0, 1), labels = c("No", "Yes"),limits = c(0,1))
fit_2_plot
```

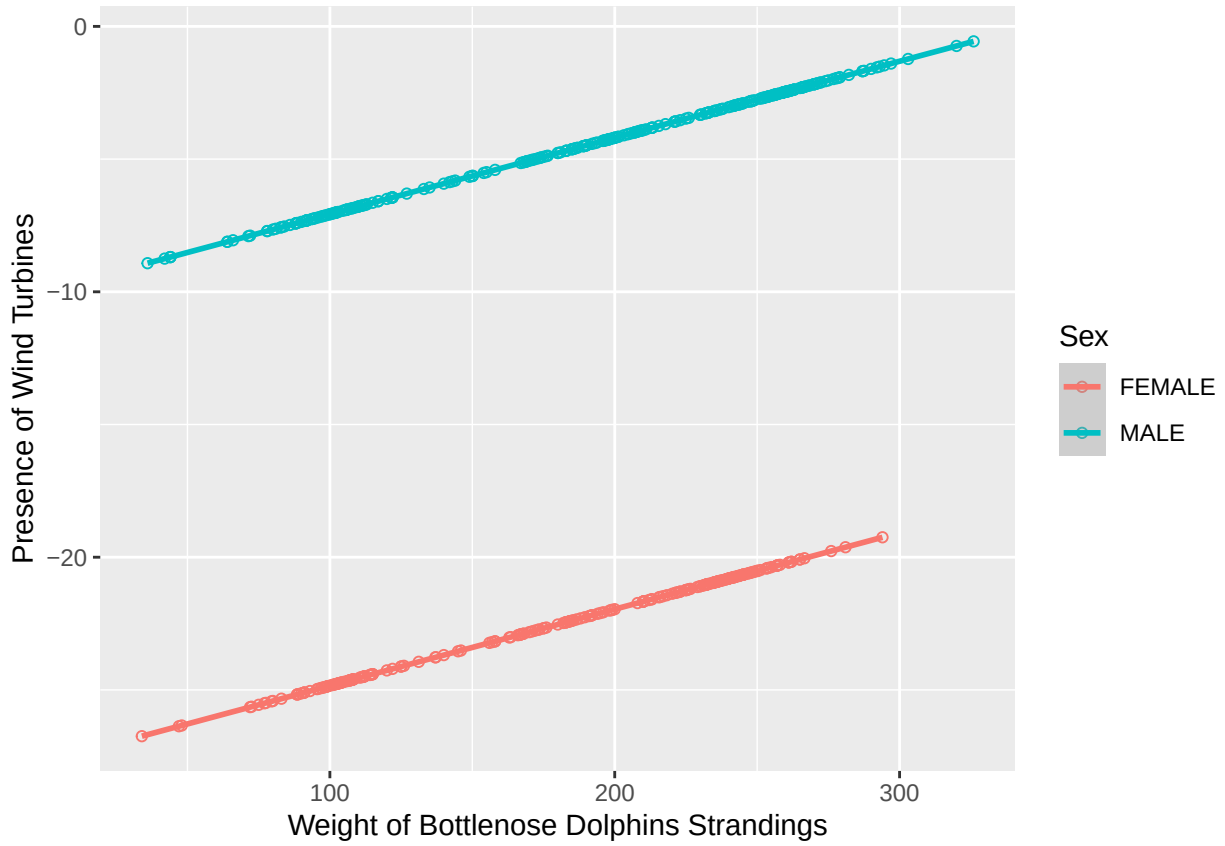



```
#Fitting a regression model (Wind Turbine presence & Weight)
fit_3 <- glm(turbine_presence~Sex+Length,family='binomial', data = cleaned_strandings)
#Summary of the regression model fit_2
summary(fit_3)
```

```
##
## Call:
## glm(formula = turbine_presence ~ Sex + Length, family = "binomial",
##     data = cleaned_strandings)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.772e+01  1.763e+03  -0.016  0.98746
## SexMALE      1.776e+01  1.763e+03   0.010  0.99196
## Length       2.882e-02  9.709e-03   2.969  0.00299 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 106.51  on 517  degrees of freedom
## Residual deviance:  77.10  on 515  degrees of freedom
## AIC: 83.1
##
## Number of Fisher Scoring iterations: 20
```

```
#Log odds of turbines presence
cleaned_strandings$log_odds_turbines<-predict(fit_3,type="link")

#Plot for fit_2
fit_3_plot <- ggplot(cleaned_strandings, aes(x =Length,y=log_odds_turbines,color=Sex)) +
geom_point(shape = 1) +
geom_smooth(method=glm)+
xlab("Weight of Bottlenose Dolphins Strandings") +
ylab("Presence of Wind Turbines")
#Plot
fit_3_plot
```

```
## `geom_smooth()` using formula = 'y ~ x'
```




```
# use to find difference in length in states with offshore wind
diff_length <- lm(Length ~ State, data = turbine_data)
#summary of linear regression
summary(diff_length)
```

```
##
## Call:
## lm(formula = Length ~ State, data = turbine_data)
##
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
## -184.672  -61.876    8.211   62.328  141.328
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   260.89      25.02  10.425 2.62e-16 ***
## StateNY       -59.98      30.95  -1.938  0.05632 .
## StateRI        42.11      79.14   0.532  0.59618
## StateVA       -76.22      27.07  -2.816  0.00619 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.07 on 76 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.08104
## F-statistic: 3.322 on 3 and 76 DF,  p-value: 0.02413
```

```
#use mutate to create state binary variables
turbine_data <- turbine_data %>%
  mutate(
    VA = ifelse(State == "VA", 1, 0),
    NY = ifelse(State == "NY", 1, 0),
    RI = ifelse(State == "RI", 1, 0),
    MA = ifelse(State == "MA", 1, 0))

#creating a linear regression model
fit_6 <- lm(Length ~ VA + NY + RI + MA, data = turbine_data)
#summary of linear regression
summary(fit_6)
```

```
##
## Call:
## lm(formula = Length ~ VA + NY + RI + MA, data = turbine_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -184.672  -61.876    8.211   62.328  141.328
##
## Coefficients: (1 not defined because of singularities)
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   260.89      25.02  10.425 2.62e-16 ***
## VA            -76.22      27.07  -2.816  0.00619 **
## NY            -59.98      30.95  -1.938  0.05632 .
## RI             42.11      79.14   0.532  0.59618
## MA                NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 75.07 on 76 degrees of freedom
## Multiple R-squared:  0.1159, Adjusted R-squared:  0.08104
## F-statistic: 3.322 on 3 and 76 DF,  p-value: 0.02413
```

```
#finding the mean of length within VA
mean_length_VA <- mean(turbine_data$Length[turbine_data$VA == "1"])
mean_length_VA
```

```
## [1] 184.6717
```

```
#finding the mean of length within VA
mean_length_NY <- mean(turbine_data$Length[turbine_data$NY == "1"])
mean_length_NY
```

```
## [1] 200.9059
```

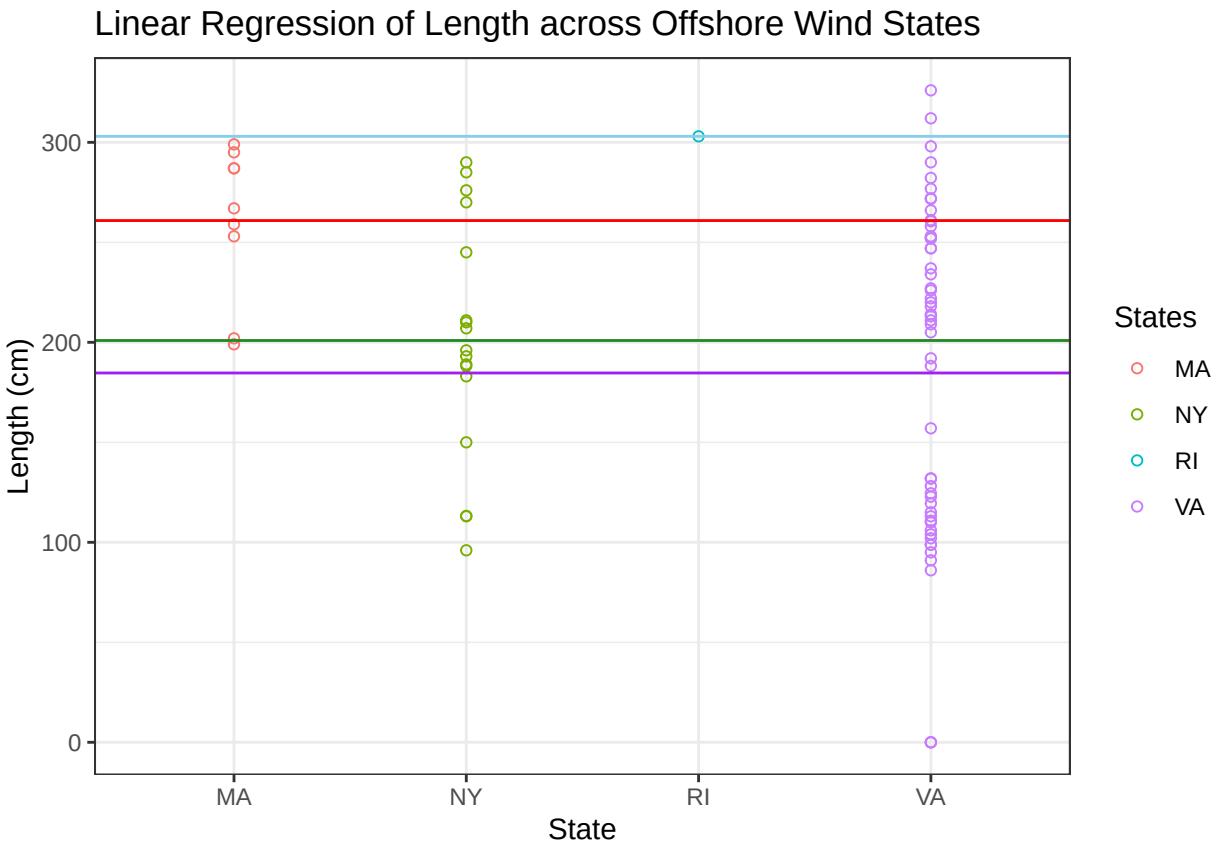
```
#finding the mean of length within VA
mean_length_RI <- mean(turbine_data$Length[turbine_data$RI == "1"])
mean_length_RI
```

```
## [1] 303
```

```
#finding the mean of length within VA
mean_length_MA <- mean(turbine_data$Length[turbine_data$MA == "1"])
mean_length_MA
```

```
## [1] 260.8889
```

```
#plot of data
plot6 <- ggplot(turbine_data, aes(x = State, y = Length, color=State)) +
    geom_point(shape = 1) +
    geom_hline(yintercept = 260.8889, color = "red") +
    geom_hline(yintercept = 200.9059, color = "forestgreen") +
    geom_hline(yintercept = 303, color = "skyblue") +
    geom_hline(yintercept = 184.6717, color = "purple") +

  xlab("State") +
  ylab("Length (cm)") +
  ggtitle("Linear Regression of Length across Offshore Wind States") +
  labs(color = "States") +
theme_bw()
plot6
```


Linear Regression of Length across Offshore Wind States
Length (cm)
300
200
100
0
MA
NY
RI
VA
State
States
MA
NY
RI
VA