EDS_assignment_5

M.Kinneen

2023-02-18

1. Environment set up

```
library(dplyr)
library(ggplot2)
library(RColorBrewer)
library(ggridges)
```

2. Data read

Read in and check data.

```r
sisco_data <- read.csv("./data/siscowet.csv", header = TRUE)

#have a look
head(sisco_data)
```

```
##       locID pnldep mesh fishID  sex age len  wgt
## 1 Deer Park  36.74  2.5  19108 <NA>  NA 316  400
## 2 Deer Park  40.09  3.0  19109 <NA>  NA 396  700
## 3 Deer Park  41.46  5.0  19110    M  NA 590 1800
## 4 Deer Park  41.46  5.0  19111    M  NA 516 1500
## 5 Deer Park  43.45  5.5  19112 <NA>  NA 414  800
## 6 Deer Park  45.58  4.0  19113    M  NA 481 1000
```

```r
summary(sisco_data)
```

```
##     locID               pnldep            mesh           fishID     
##  Length:780         Min.   : 15.40   Min.   :2.000   Min.   :19108  
##  Class :character   1st Qu.: 45.20   1st Qu.:2.500   1st Qu.:19362  
##  Mode  :character   Median : 59.60   Median :3.500   Median :19558  
##                     Mean   : 56.23   Mean   :3.576   Mean   :19576  
##                     3rd Qu.: 69.05   3rd Qu.:4.500   3rd Qu.:19816  
##                     Max.   :108.69   Max.   :6.000   Max.   :20053  
##                                                                     
##      sex                 age             len             wgt       
##  Length:780         Min.   : 7.00   Min.   :240.0   Min.   :  150  
##  Class :character   1st Qu.:10.00   1st Qu.:443.0   1st Qu.:  775  
##  Mode  :character   Median :11.00   Median :493.0   Median : 1100  
##                     Mean   :11.45   Mean   :487.1   Mean   : 1175  
##                     3rd Qu.:12.25   3rd Qu.:536.2   3rd Qu.: 1500  
##                     Max.   :21.00   Max.   :762.0   Max.   :15800  
##                     NA's   :580                     NA's   :1
```

```r
#check for NA
sapply(sisco_data, function(x) sum(is.na(x)))
```

```
##  locID pnldep   mesh fishID    sex    age    len    wgt 
##      0      0      0      0     59    580      0      1
```

Sex and age columns have large amounts of NA values. Also one in weight.

3. Initial exploratory plots

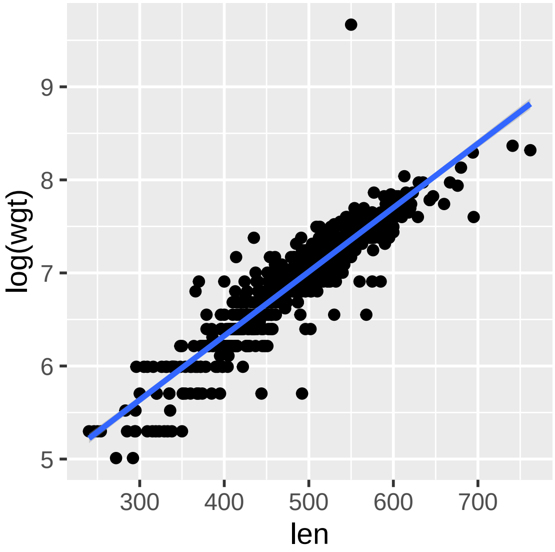
```
univar_exploratory <- function(df){
  columns<-c(colnames(df)) #create list of columns
  for(i in 1:length(columns)){   #Loop through plotting hist of each
  var<- df[,i]   #select var for plotting
  title<-paste(columns[i])
  if(is.numeric(var)==TRUE){   #if var is numeric, plot, otherwise pass
    hist(var,
         main = title)
  }else{
    print(paste0(title,"(column ",i,") ", "is a non-numeric column"))
  }
  }
}

#length_weight plot to check outliers
ggplot(sisco_data,aes(x=len,y = log(wgt)))+
  geom_point()+
  stat_smooth(method = "lm")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```

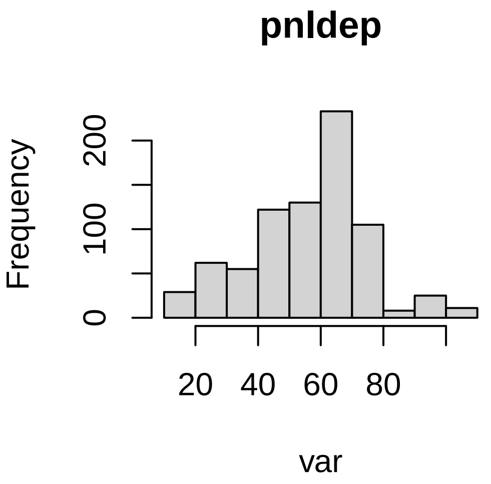
```
## Warning: Removed 1 rows containing non-finite values (‘stat_smooth()‘).
```

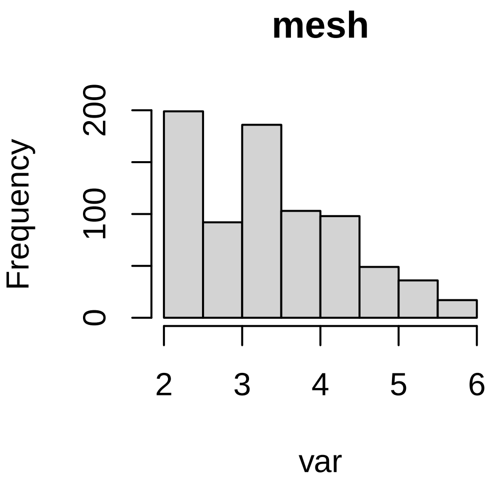
```
## Warning: Removed 1 rows containing missing values (‘geom_point()‘).
```

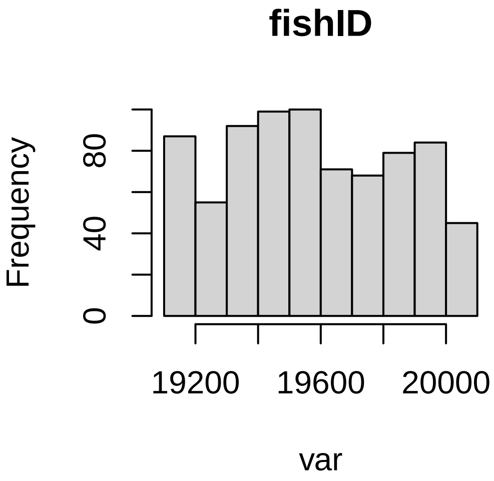


```
univar_exploratory(df = sisco_data)
```

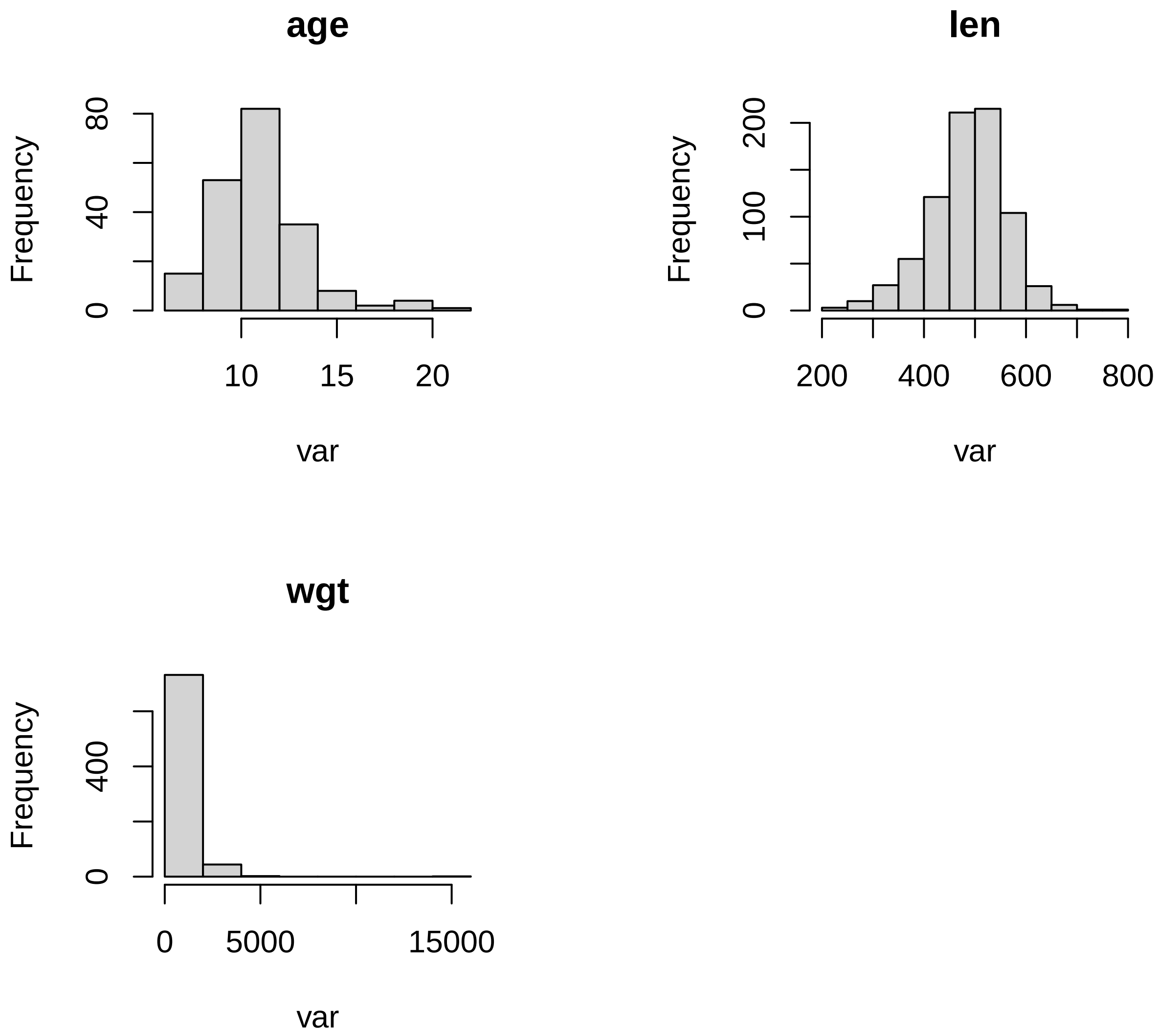
```
## [1] "locID(column 1) is a non-numeric column"
```







```
## [1] "sex(column 5) is a non-numeric column"
```



PNLdepth is approximately normal, median ~ 70.Mesh has right skew, median 2 - 3. Possibly log transform? Fish ID approxiamtely uniform, irrelevant for analysis.Age has right skew, median age of ~ 11. Length is normal, median of 500. Weight is heavy right skew, lagre max value likely outlier (confirmed by length-weight plot). filtering needed

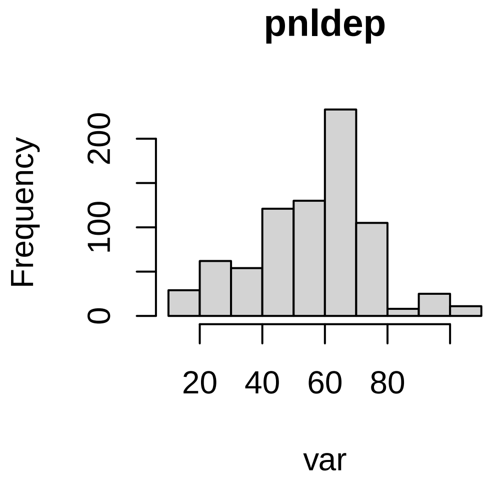
4. Clean and filter data

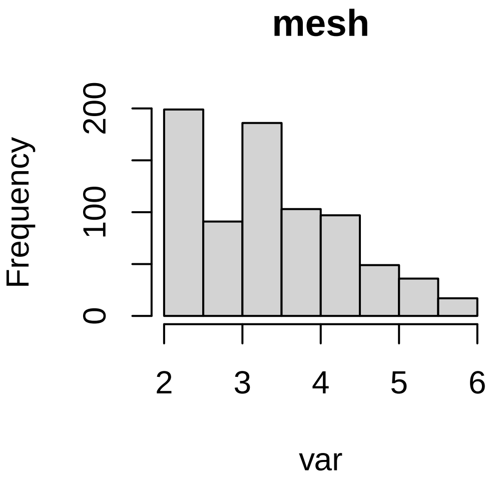
Weight less skewed though still non-normal. Do not do more filtering as larger fish will be underrepresented in data.

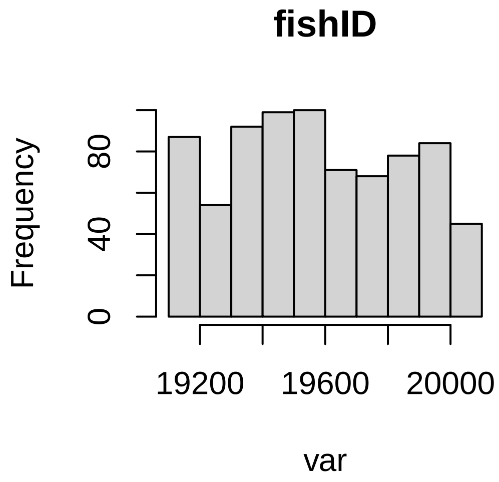
```
sisco_data_cleaned <- sisco_data
sisco_data_cleaned<- sisco_data%>%
  filter(wgt <= 7000)%>% #filter max value from wgt
  mutate(mesh_log = log(mesh)) #log transform mesh
```

```
#Repeat plots
univar_exploratory(df = sisco_data_cleaned)
```

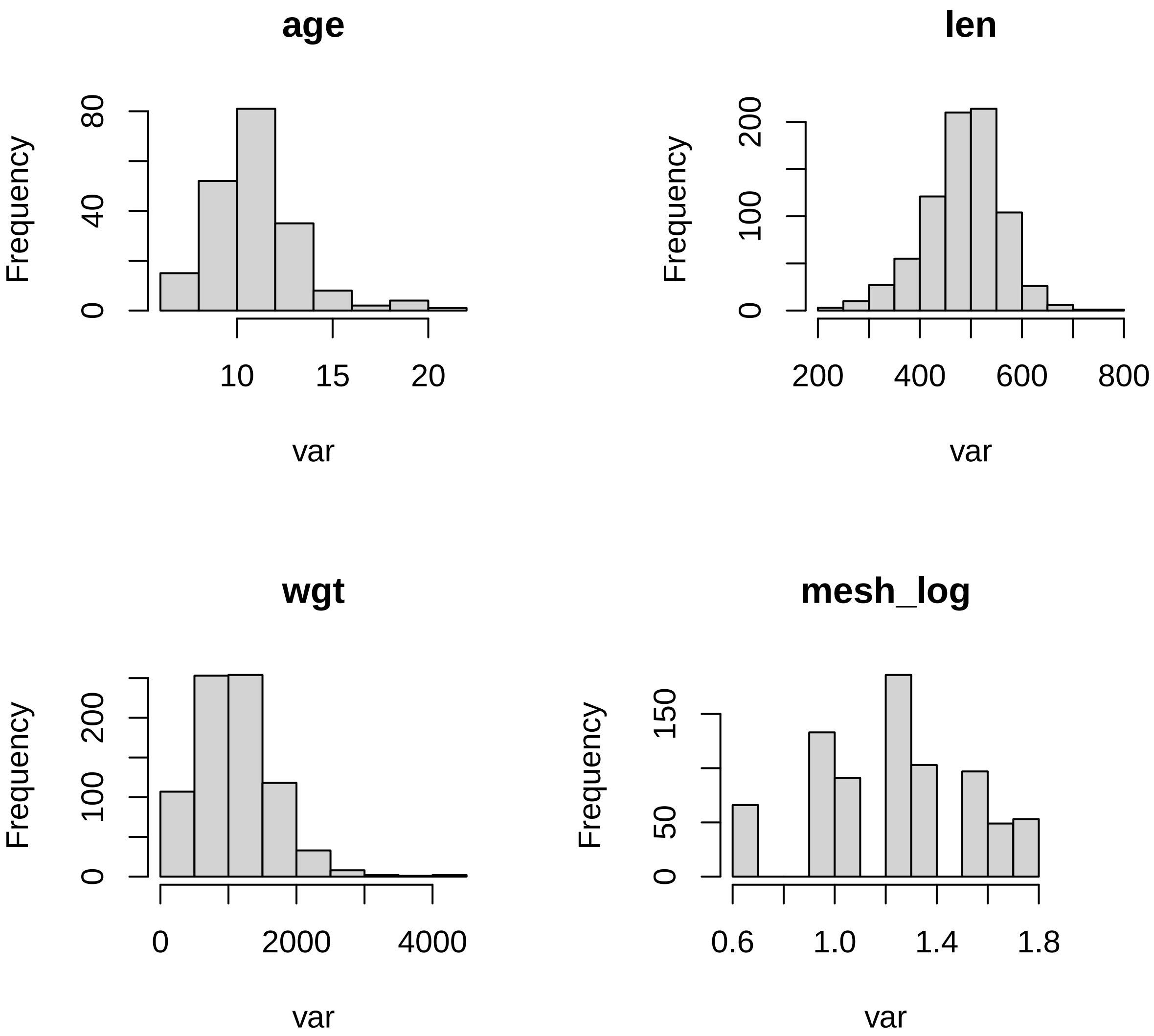
```
## [1] "locID(column 1) is a non-numeric column"
```







```
## [1] "sex(column 5) is a non-numeric column"
```


age
Frequency
0
40
80
10
15
20
var
len
Frequency
0
100
200
200
400
600
800
var
wgt
Frequency
0
100
200
0
2000
4000
var
mesh_log
Frequency
0
50
150
0.6
1.0
1.4
1.8
var

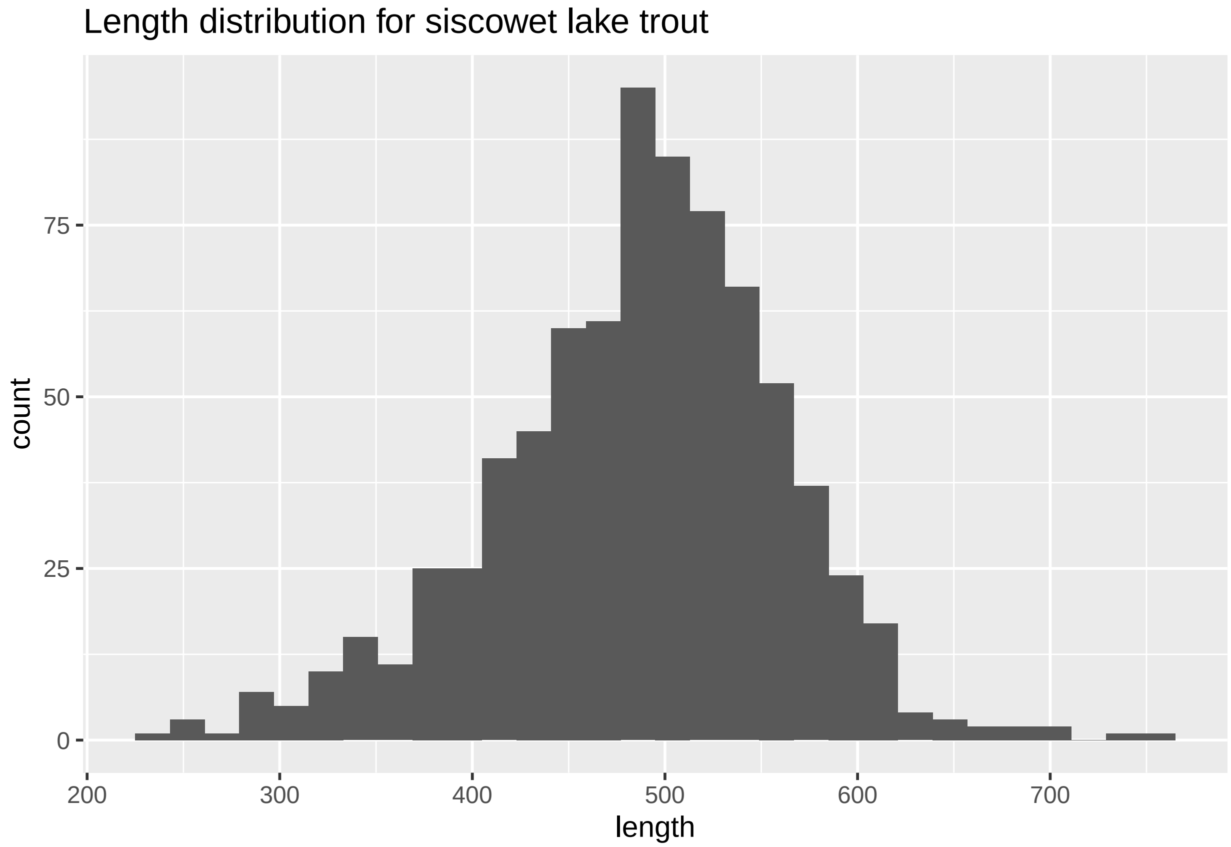
5. Exploratory plot

Plot showing the length distribution of fish across sites.

```
exploratory<-ggplot(sisco_data_cleaned,aes(x=len))+
  geom_histogram()+
  xlab("length")+
  labs(title = "Length distribution for siscowet lake trout")

exploratory
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggsave("./images/exploratory.jpg", width = 3.25, height = 2.25) #save image
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

```
#save image
```

6. Expository plot

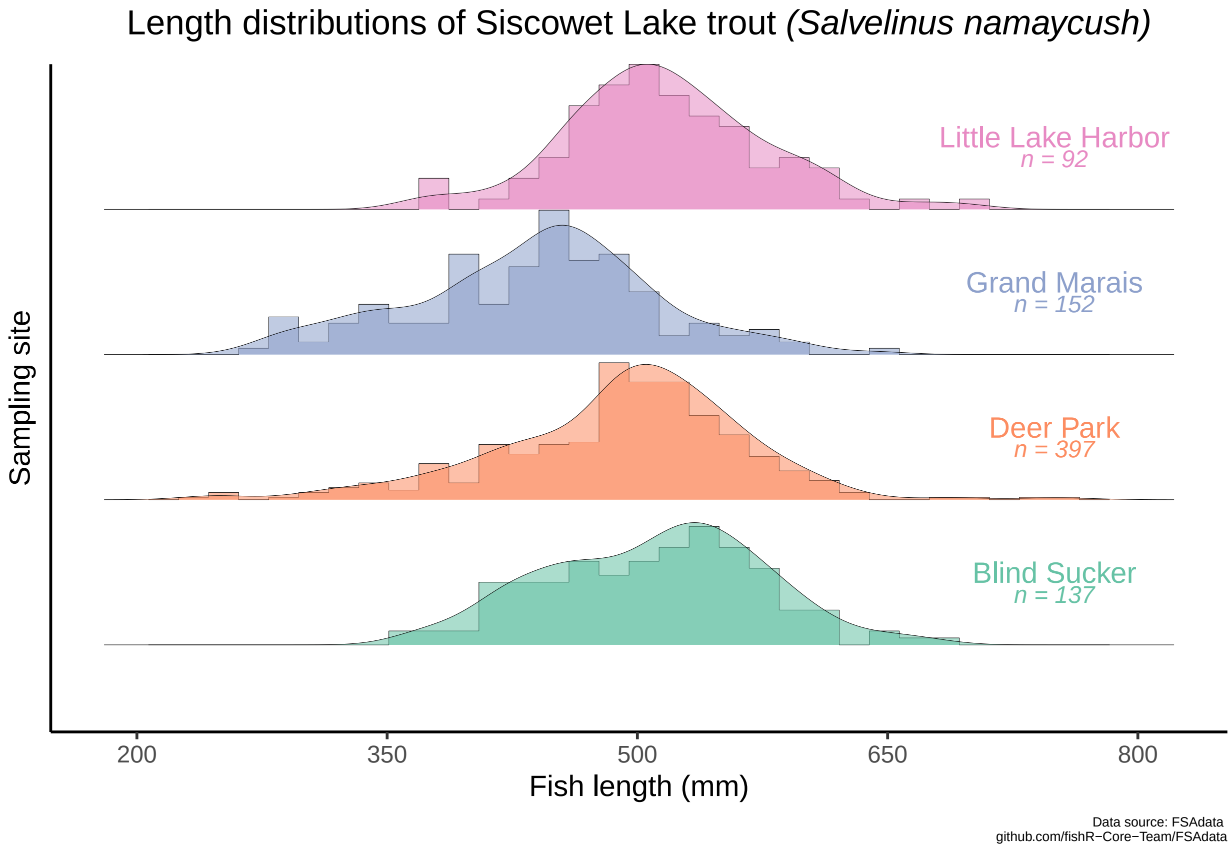
Fish length distributions across sites.

```
palette<- brewer.pal(4,"Set2") #colorblinf friendly paeltte
site_names<- unique(sisco_data_cleaned$locID) #site names for labels
samples<-table(sisco_data_cleaned$locID) # count data for lavbels

expository_plot<-
  #Define data
  ggplot(sisco_data_cleaned,aes(x = len, y = locID, alpha =0.05, fill = locID))+
  #Define geometry (histograms)
  geom_density_ridges(stat = "binline", scale = 1, size = 0.01)+
  #Overlay curve
  geom_density_ridges(scale = 1, size = 0.01)+
    scale_fill_manual(values = palette)+
  #Set axis breaks
  scale_x_continuous(breaks = seq(200,800,150))+
  #Add title
  labs(title = substitute(paste("Length distributions of Siscowet Lake trout ",
                     italic("(Salvelinus namaycush)"))),
       #Data source caption
       caption = paste("Data source: FSAdata","\n",
                        "github.com/fishR-Core-Team/FSAdata"))+
  xlab("Fish length (mm)")+
  ylab("Sampling site")+
  #Simple theme
  theme_classic()+
  #Custom theme - remove y axis title, adjust title position
  theme(
    axis.text.y  = element_blank(),
    legend.position = "none",
    axis.ticks.y = element_blank(),
    plot.title = element_text(hjust = 0.5),
    plot.subtitle = element_text(hjust = 0.5, size = 7),
    plot.caption = element_text(size = 5)
  )+
  #Add annotations for site name and number of samples.
  annotate("text",x = 750, y = 1.5, label = site_names[2],colour=palette[1])+
  annotate("text",x = 750, y = 1.35, label = paste0("n = ",samples[1])
           ,colour=palette[1],fontface = 3,size = 3)+
  annotate("text",x = 750, y = 2.5, label = site_names[1],colour=palette[2])+
   annotate("text",x = 750, y = 2.35, label = paste0("n = ",samples[2])
           ,colour=palette[2],fontface = 3,size = 3)+
  annotate("text",x = 750, y = 3.5, label = site_names[4],colour=palette[3])+
   annotate("text",x = 750, y = 3.35, label = paste0("n = ",samples[3])
           ,colour=palette[3],fontface = 3,size = 3)+
  annotate("text",x = 750, y = 4.5, label = site_names[3],colour=palette[4])+
   annotate("text",x = 750, y = 4.35, label = paste0("n = ",samples[4])
           ,colour=palette[4],fontface = 3,size = 3)
```

```
expository_plot
```

```
## ‘stat_binline()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## Picking joint bandwidth of 19.9
```



```
ggsave("./images/expository.jpg", width = 3.25, height = 2.25)
```

```
## ‘stat_binline()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## Picking joint bandwidth of 19.9
```