CHIS ASSP Mistnetting Database QA/QC

Amelia DuVall (ajduvall@uw.edu) & Emma Kelsey (ekelsey@usgs.gov)

2 April 2020

This is v.2020-04-16

This document explains the QA/QC conducted on the Channel Islands National Park (CHIS) Ashy Storm-Petrel (ASSP) Mistnetting database. The purpose of this exercise is to capture incongruent values in the database that likely arose from data entry and/or data manipulation errors. We cannot verify observor errors (e.g., if errors were made while taking morphometric measurements but were recorded properly). Whenever possible, we cross-referenced raw data (e.g., field notebooks or scanned data sheets) to verify and/or fix incongruent values.

The database is an Excel file that contains (6) sheets: 6 sheet: Banding, Banding_Data_Dictionary, CPUE, CPUE_Data_Dictionary, Mistnetting_Locations, Participant_Initials. This document is divided into two categories: Banding QAQC and CPUE QAQC. Amelia DuVall conducted QAQC on the Banding data and Emma Kelsey conducted QAQC on the CPUE metadata.

Load libraries

```
library(tidyverse)
library(lubridate)
library(ggplot2)
library(openxlsx)
library(readxl)
```

Read-in data and set-up for analysis

```
banding <- read_excel("CHIS_ASSP_mistnet_database_04162020.xlsx", sheet = "Banding",
                      col_names = TRUE, na = c("NA", "ND"))

cpue <- read_excel("CHIS_ASSP_mistnet_database_04162020.xlsx", sheet = "CPUE",
                   col_names = TRUE, na = c("NA", "ND"))
```

Banding Data QAQC

Filter data to ASSP species and banded individuals only.

```
ASSP <- group_by(.data = banding) %>%
          filter(species == "ASSP") %>%
          filter(band_no != "notbanded") %>%
          ungroup()
```

Summarize band numbers and capture rates.

Summarize data by band number and determine capture rate for each band number (e.g., each individual).

```
summary <-  group_by(.data = ASSP, band_no) %>%
              summarise(no_captures = n()) %>%
              ungroup()
```

There are 3643 unique band numbers.

Summarize capture rates.

```
summarycaptures <- group_by(summary, no_captures) %>%
                      summarise(count = n()) %>%
                      ungroup()

show(summarycaptures)
```

```
## # A tibble: 4 x 2
##   no_captures count
##         <int> <int>
## 1           1  3465
## 2           2   159
## 3           3    16
## 4           4     3
```

Summarize recapture rates.

```
recap <- group_by(.data = ASSP, recapture) %>%
            summarise(no_captures = n()) %>%
            ungroup()

show(recap)
```

```
## # A tibble: 3 x 2
##   recapture no_captures
##   <chr>           <int>
## 1 N                3606
## 2 SNR                44
## 3 Y                 193
```

The recapture rates will not necessarily match the unique band numbers because we encounteed some individuals only once as a recapture (i.e., we did not band them). These bands need to be cross-referenced against BBL data to determine when/where they were first banded. This database does not contain all ASSP mistnetting banding records from CHIS.

Another way to check the recapture field is to sort the band numbers sequentially and look for outliers. In theory, these would be on someone else's banding permit (and a recapture).

```
seq <- arrange(.data = ASSP, band_no)
```

We visually inspected these data and flagged any bands that were out of order to cross-reference against raw data.

Capture time

```
# Change data type of time stamp and isolate hour as a new field.
ASSP$capture_time <- mdy_hm(ASSP$capture_time, tz="US/Pacific")
```

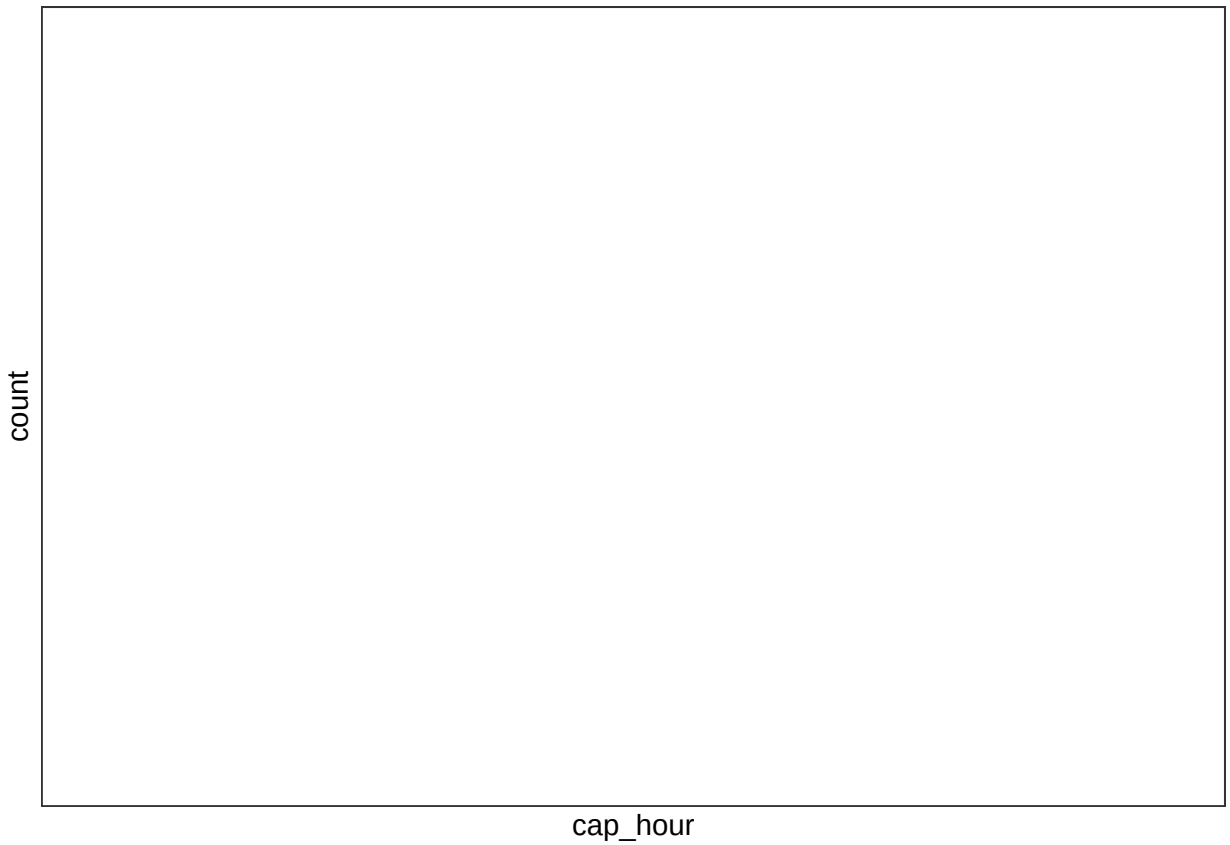
```
## Warning: All formats failed to parse. No formats found.
```

```
ASSP$cap_hour <- hour(ASSP$capture_time)

# Which unique hour values are represented in the data?
unique(ASSP$cap_hour)
```

```
## [1] NA
```

```
# The unique values at 11, 10, and 12 are incongruent with nighttime mistnetting. Perhaps these were no
# Also, 4 also seems like a late capture time, but it's plausible.

# Create histogram of capture time.
ggplot(data = ASSP) +
  geom_histogram(mapping = aes(x = cap_hour), binwidth = 1) +
  theme_bw()
```

```
## Warning: Removed 3843 rows containing non-finite values (stat_bin).
```




```
# Isolate questionable data.
cap_hour_chk <- group_by(.data = ASSP, cap_hour) %>%
                  filter(cap_hour %in% c(4, 10, 11, 12)) %>%
                  ungroup()

# Export to csv and cross-reference raw data.
#write.csv(cap_hour_chk, "captimeQAQC.csv")
```

Release time

```
# Change data type of release time stamp and isolate hour as new field.
ASSP$release_time <- mdy_hm(ASSP$release_time, tz="US/Pacific")
```

```
## Warning: All formats failed to parse. No formats found.
```

```
ASSP$rel_hour <- hour(ASSP$release_time)

# Which unique hour values are represented in the data?
unique(ASSP$rel_hour)
```

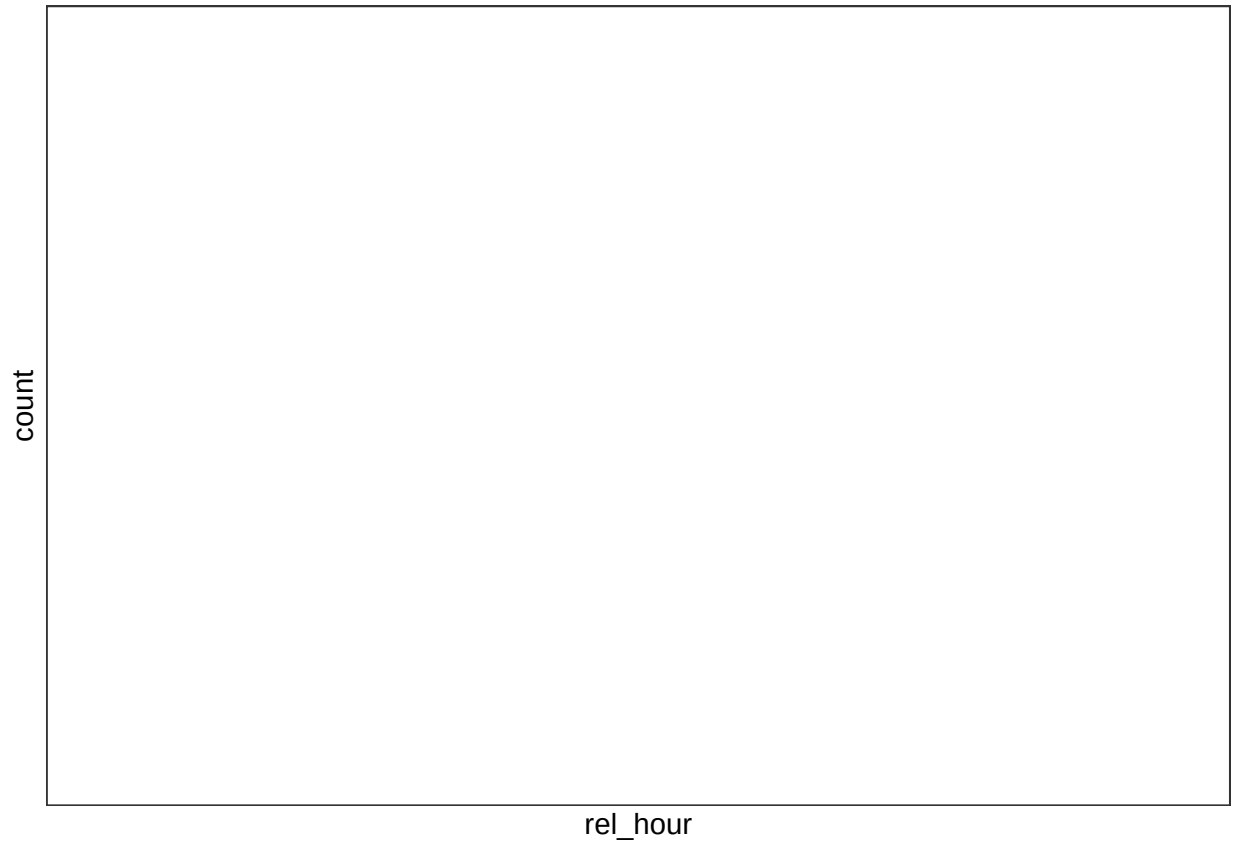
```
## [1] NA
```

```
# The unique values at 12 are incongruent with nighttime mistnetting. Perhaps these were not entered in
# Double-check 3am and 4am also.

summary(ASSP$rel_hour)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##      NA      NA      NA     NaN      NA      NA    3843
```

```
# There are a lot of NA's since the release time was often unrecorded.

ggplot(data = ASSP) +
  geom_histogram(mapping = aes(x = rel_hour), binwidth = 1) +
  theme_bw()
```

```
## Warning: Removed 3843 rows containing non-finite values (stat_bin).
```




```
#Isolate questionable data
rel_hour_chk <- group_by(.data = ASSP, rel_hour) %>%
                  filter(rel_hour %in% c(3, 4, 12)) %>%
                  ungroup()

# Export to csv and cross-reference raw data.
#write.csv(rel_hour_chk, "reltimeQAQC.csv")
```

Brood patch

```
# What are the unique BP values represented in the data?
unique(ASSP$BP)
```

```
##  [1] "4.5" "1"   "1.5" "3"   "5"   "2"   "4"   NA    "0"   "B"   "PD"  "D"
## [13] "b"
```

```
# All these values fall within the range of values included in data dictionary.
```

Mass

There are three mass values recorded in the database: mass (uncorrected), mass (tare), and mass (corrected). The mass (corrected) field is the remainder of mass (uncorrected) - mass (tare).

```
unique(ASSP$mass_corr)
```

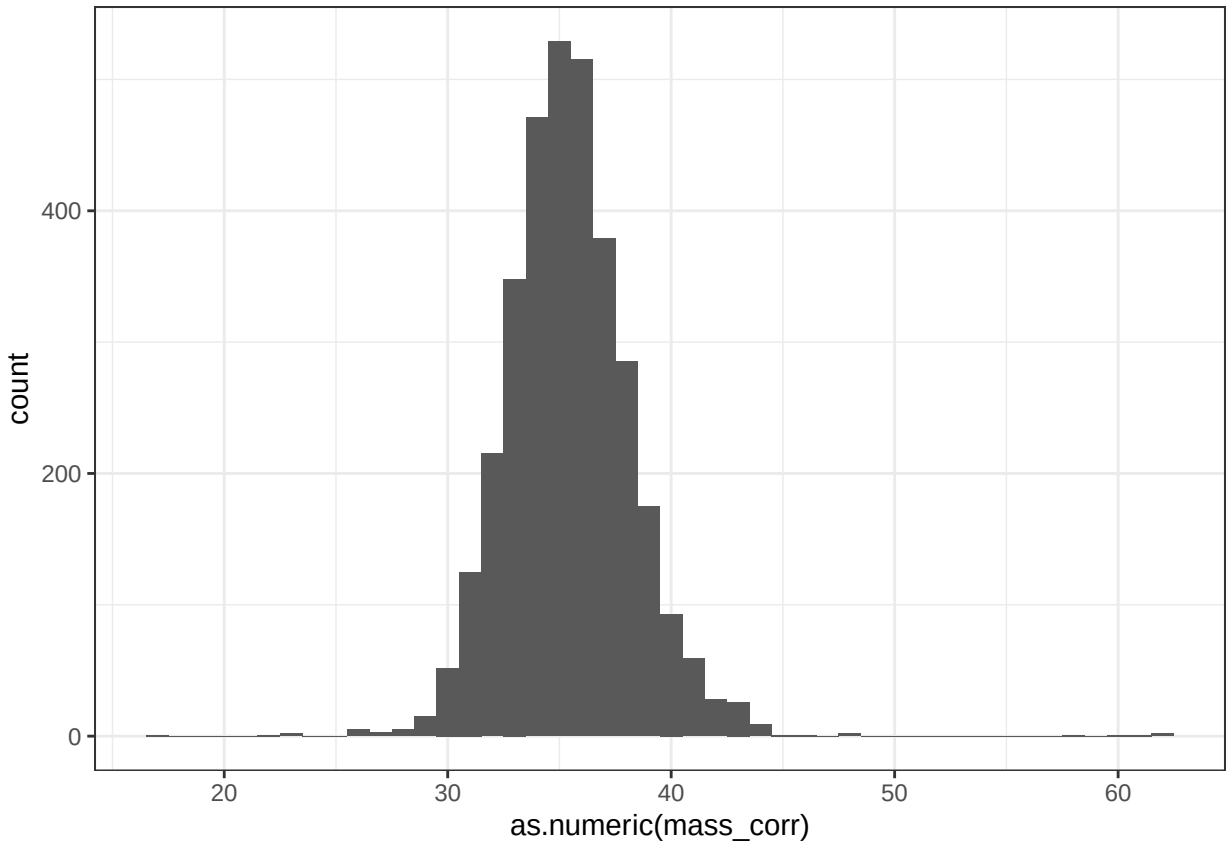
```
##  [1] 38 30 37 39 33 34 40 23 32 35 NA 36 31 42 41 43 62 60 58 61 26 44 17 29 48
## [26] 27 28 45 46 22
```

```
summary(ASSP$mass_corr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17.00   34.00   35.00   35.44   37.00   62.00     493
```

```
# Large range in values recorded.

ggplot(data = ASSP) +
  geom_histogram(mapping = aes(x = as.numeric(mass_corr)), binwidth = 1) +
  theme_bw()
```

```
## Warning: Removed 493 rows containing non-finite values (stat_bin).
```




```
# Mass (g) values reported in Adams (2016) paper: 36.1 +/- 2.8 (female) and 34.7 +/- 2.1 (male).

# Isolate questionable data.
mass_corr_chk <- filter(ASSP, mass_corr < 25 | mass_corr > 50)

# Export to csv and cross-reference raw data.
#write.csv(mass_corr_chk, "masscorrQAQC.csv")
```

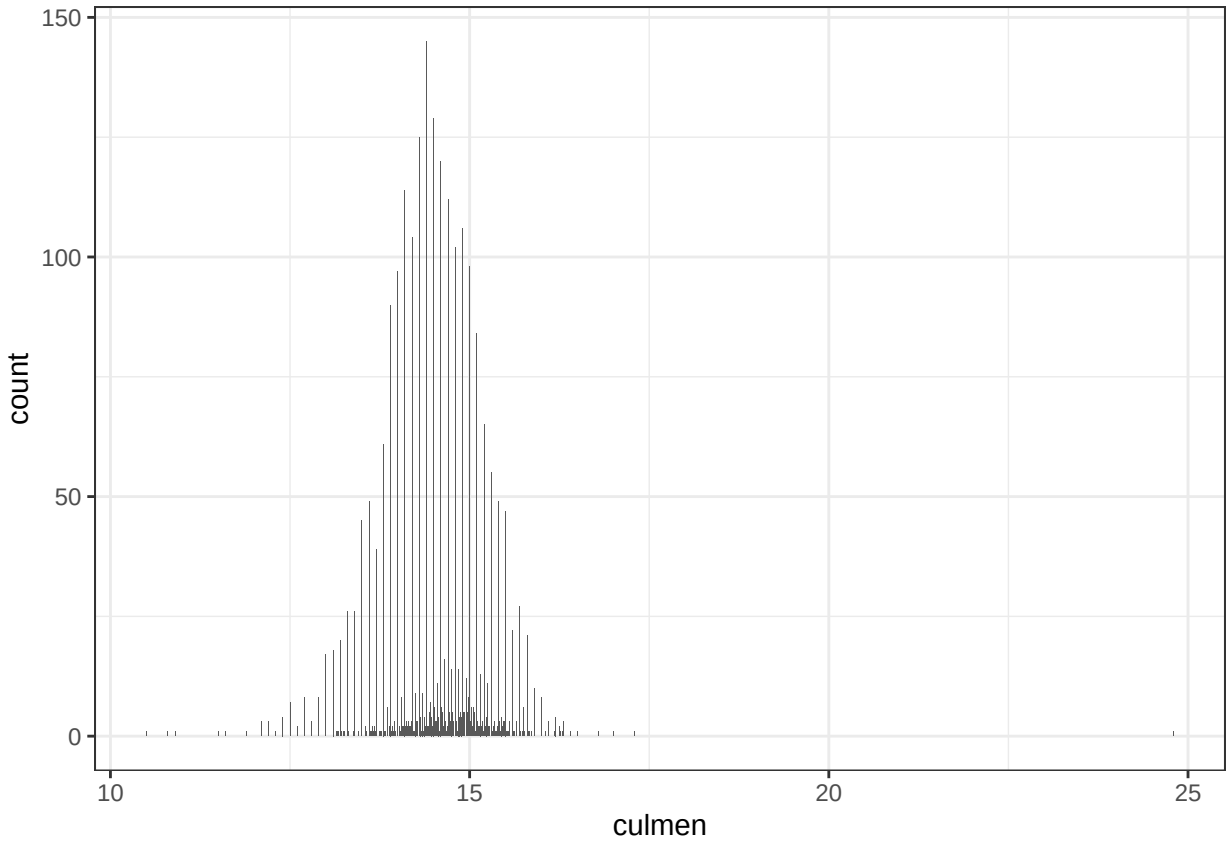
Culmen

```
summary(ASSP$culmen)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   10.50   14.10   14.50   14.51   14.99   24.80    1252
```

```
# Large range in values recorded.

ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = culmen)) +
  theme_bw()
```

```
## Warning: Removed 1252 rows containing non-finite values (stat_count).
```




```
# Bill length (mm) values reported in Adams (2016) paper: 14.9 +/- 0.5 (f) and 14.6 +/- 0.8 (m).
# Values reported in Pyle guide: 13.1-15.2 (95% CI).

#Isolate questionable data
culmen_chk <- filter(ASSP, culmen < 13 | culmen > 16)

# Export to csv and cross-reference raw data.
#write.csv(culmen_chk, "culmenQAQC.csv")
# Update (4/15): It seems like the range I picked (13-16) was too narrow. Most were not typos.
```

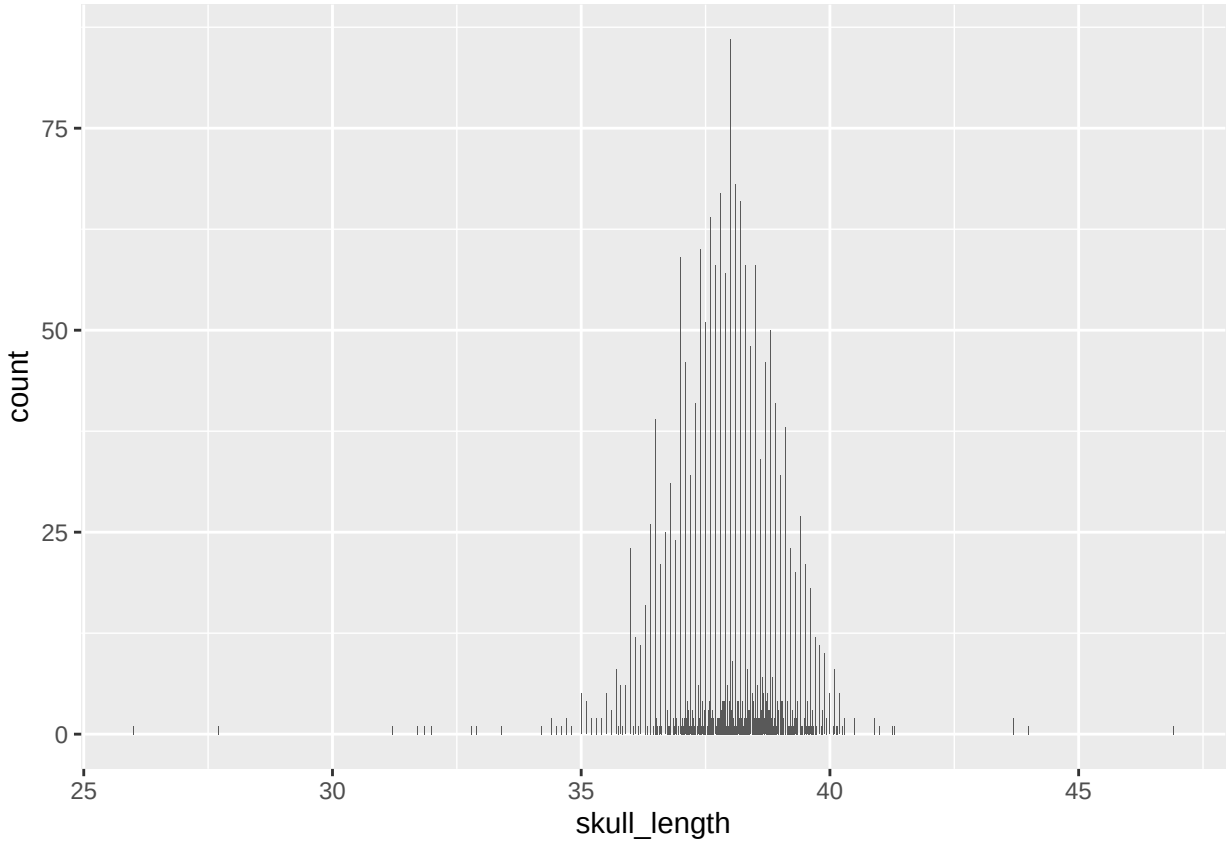
Skull length

```
summary(ASSP$skull_length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   26.00   37.30   38.00   37.93   38.65   46.90    1804
```

```
ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = skull_length))
```

```
## Warning: Removed 1804 rows containing non-finite values (stat_count).
```




```
# Skull length (mm) values reported in Adams (2016): 38.1 +/- 1.1 (f) and 37.9 +/- 0.8 (m).
# No information on skull length in Pyle guide.

#Isolate questionable data.
skull_chk <- filter(ASSP, skull_length < 35 | skull_length > 41)

# Export to csv and cross-reference raw data.
#write.csv(skull_chk, "skullQAQC.csv")
```

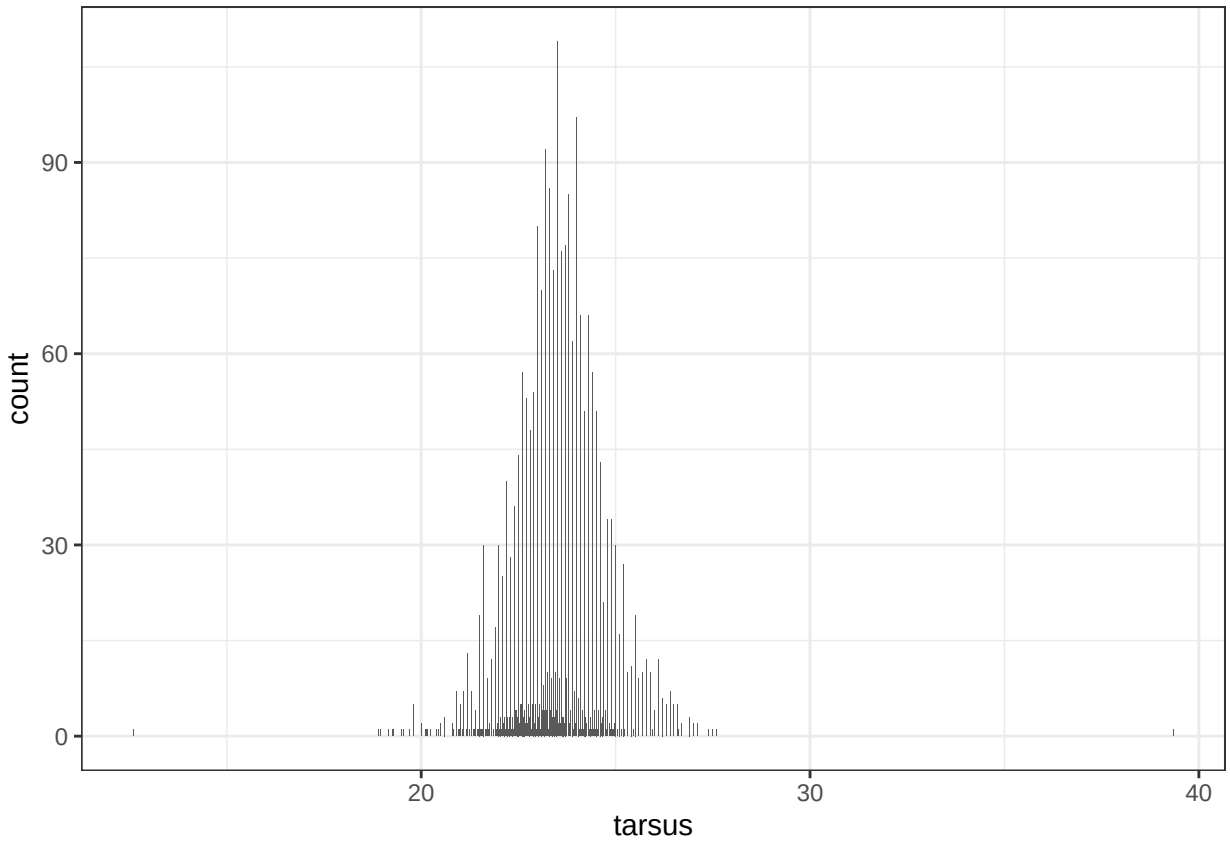
Tarsus

```
summary(ASSP$tarsus)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   12.60   22.80   23.45   23.47   24.10   39.34    1251
```

```
# Large range in values recorded.

ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = tarsus)) +
  theme_bw()
```

```
## Warning: Removed 1251 rows containing non-finite values (stat_count).
```




```
# Tarsus (mm) values reported in Adams (2016): 23.2 +/- 0.9 (f) and 23.1 +/- 0.8 (m).
# Values reported in Pyle guide: 21-25 (95% CI).

#Isolate questionable data.
tarsus_chk <- filter(ASSP, tarsus < 19 | tarsus > 27)

# Export to csv and cross-reference raw data.
#write.csv(tarsus_chk, "tarsusQAQC.csv")
```

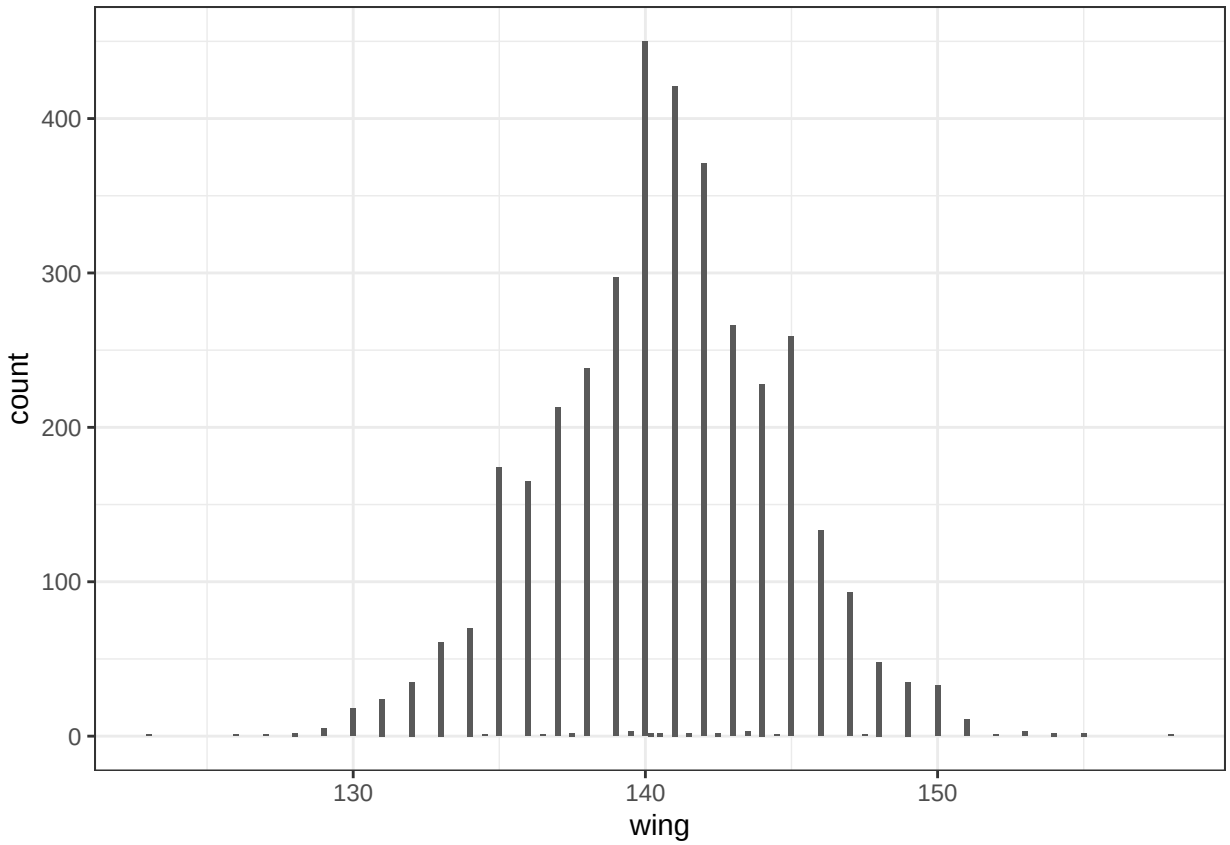
Wing chord

```
summary(ASSP$wing)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   123.0   138.0   141.0   140.6   143.0   158.0     161
```

```
# Large range in values recorded.

ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = wing)) +
  theme_bw()
```

```
## Warning: Removed 161 rows containing non-finite values (stat_count).
```

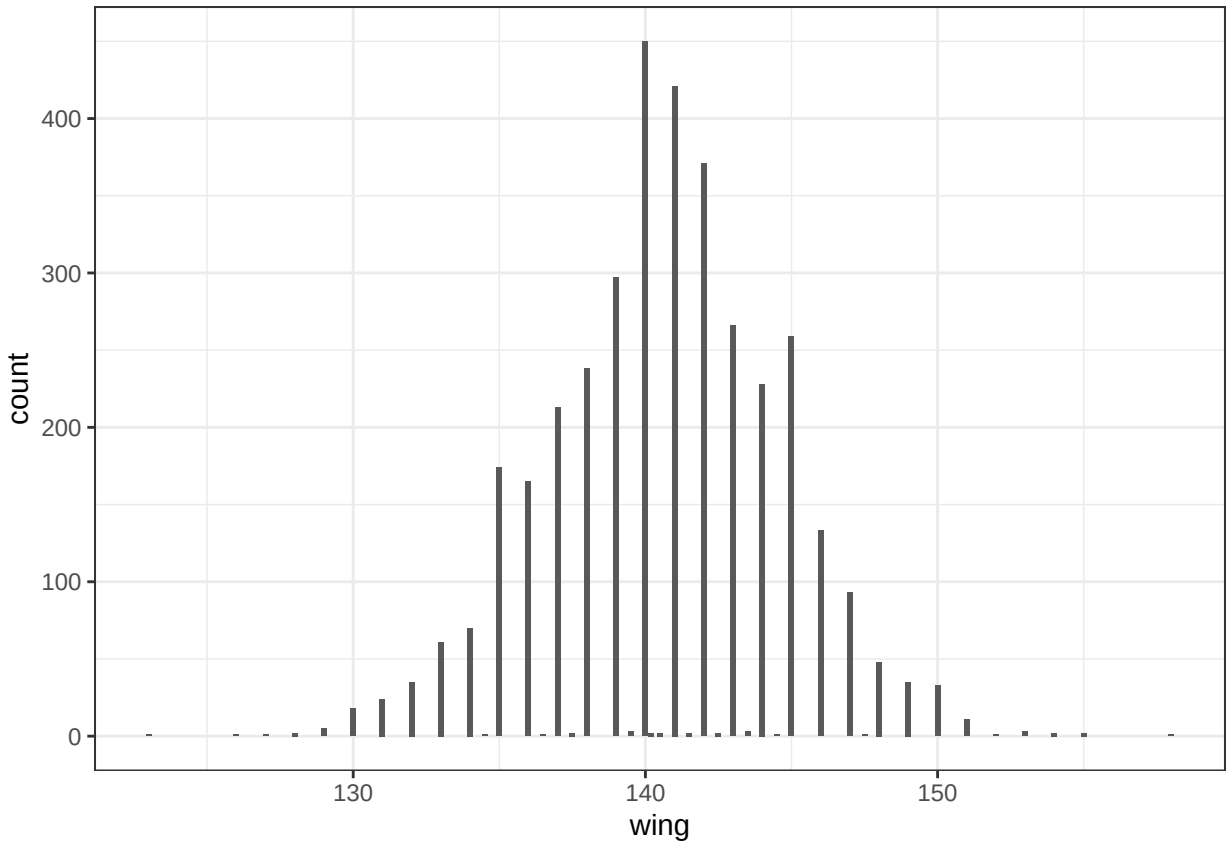



```
# Max flat wing (mm) values reported in Adams (2016): 142.7 +/- 2.8 (f) and 140.4 +/- 3.3 (m).
# Values reported in Pyle guide: 132-148 (95% CI).

#Isolate questionable data.
wing_chk <- filter(ASSP, wing < 130 | wing > 150)

# Export to csv and cross-reference raw data.
#write.csv(wing_chk, "wingQAQC.csv")
```

We decided to check for a bimodal distribution of the wing chord morphometric data once the data entry errors had been rectified. A bimodal distribution could be an indication of different methods used to measure wing chord (e.g, flattened wing chord versus relaxed wing chord).

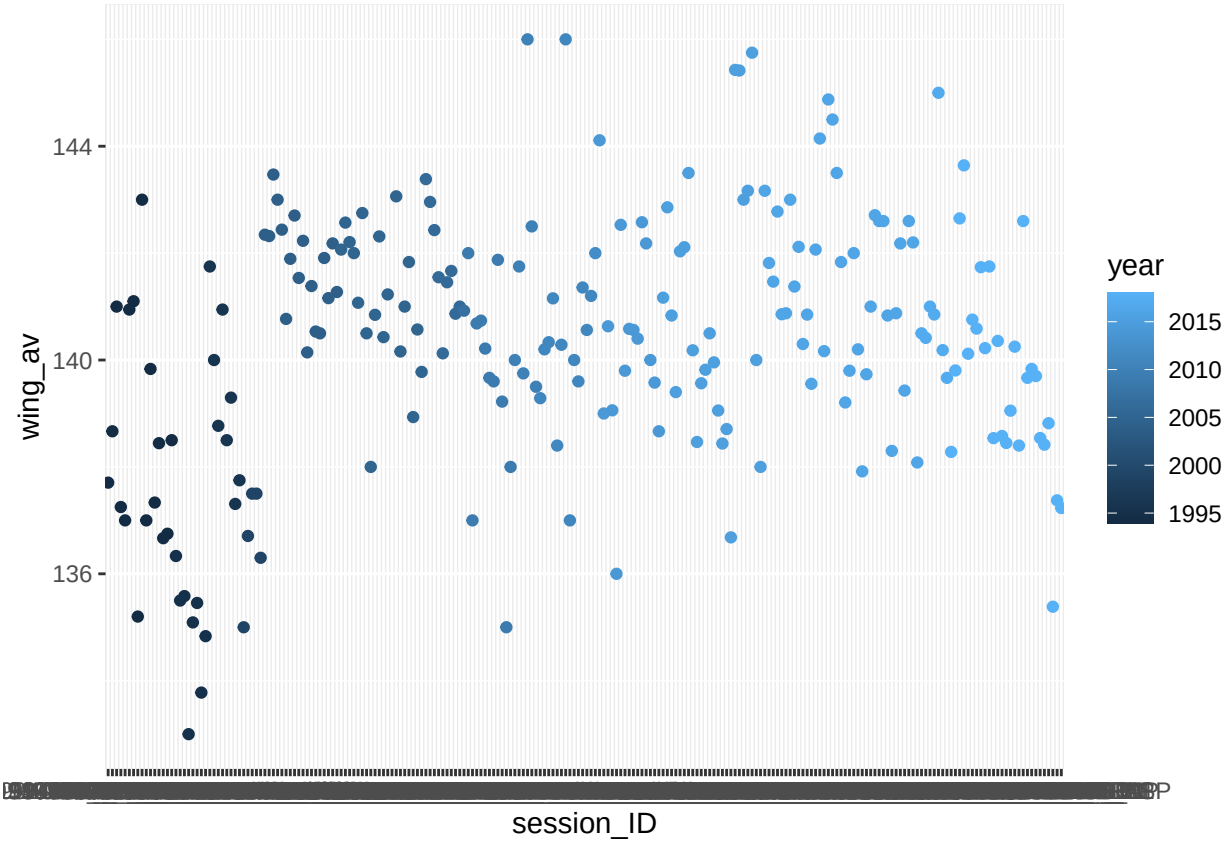
```
# Plot wing chord values.
ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = wing)) +
  theme_bw()
```

```
## Warning: Removed 161 rows containing non-finite values (stat_count).
```

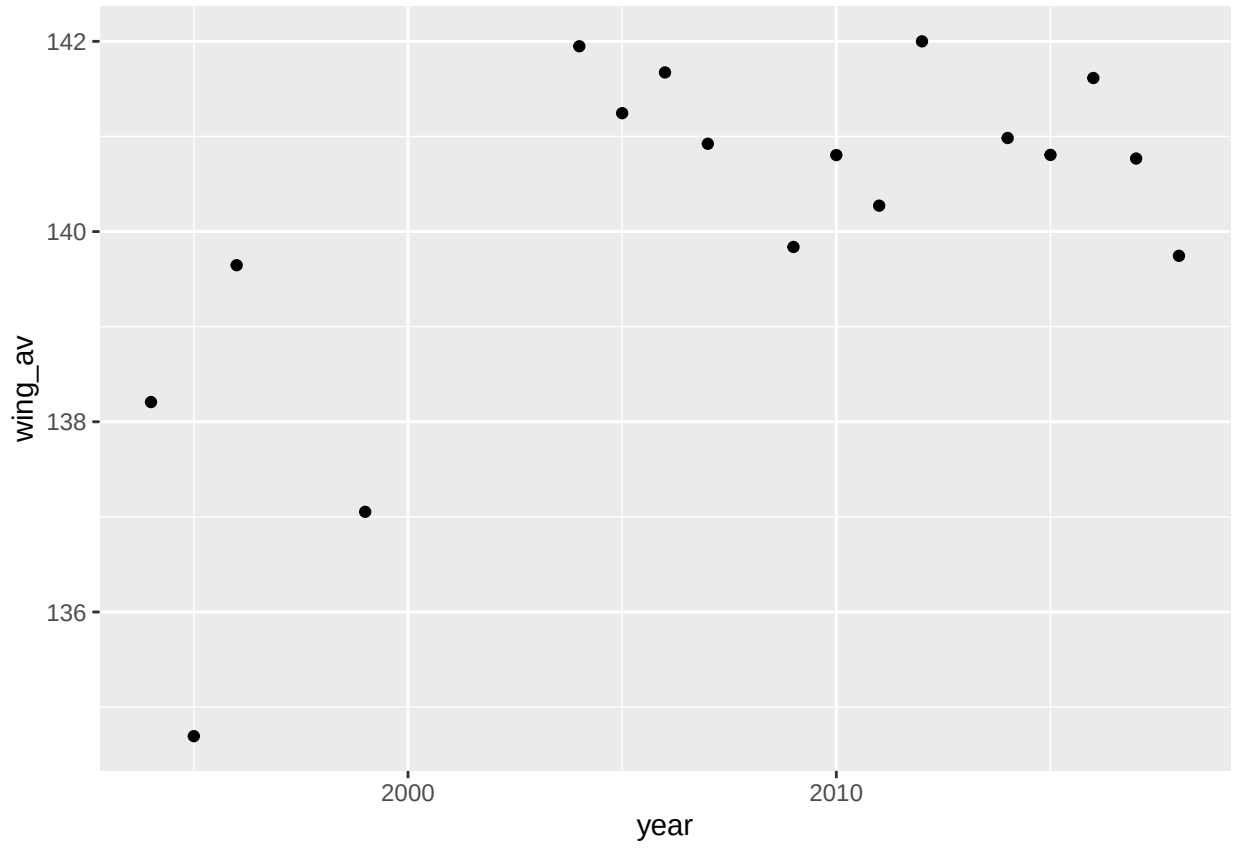



```
# There does not appear to be a bimodal distribution.

# How about mean wing chord observed per session?
WC_ses <- group_by(.data = ASSP, session_ID, year) %>%
          summarise(wing_av = mean(wing, na.rm = TRUE)) %>%
          ungroup()

ggplot(WC_ses) +
  geom_point(aes(x = session_ID, y = wing_av, color = year))
```

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

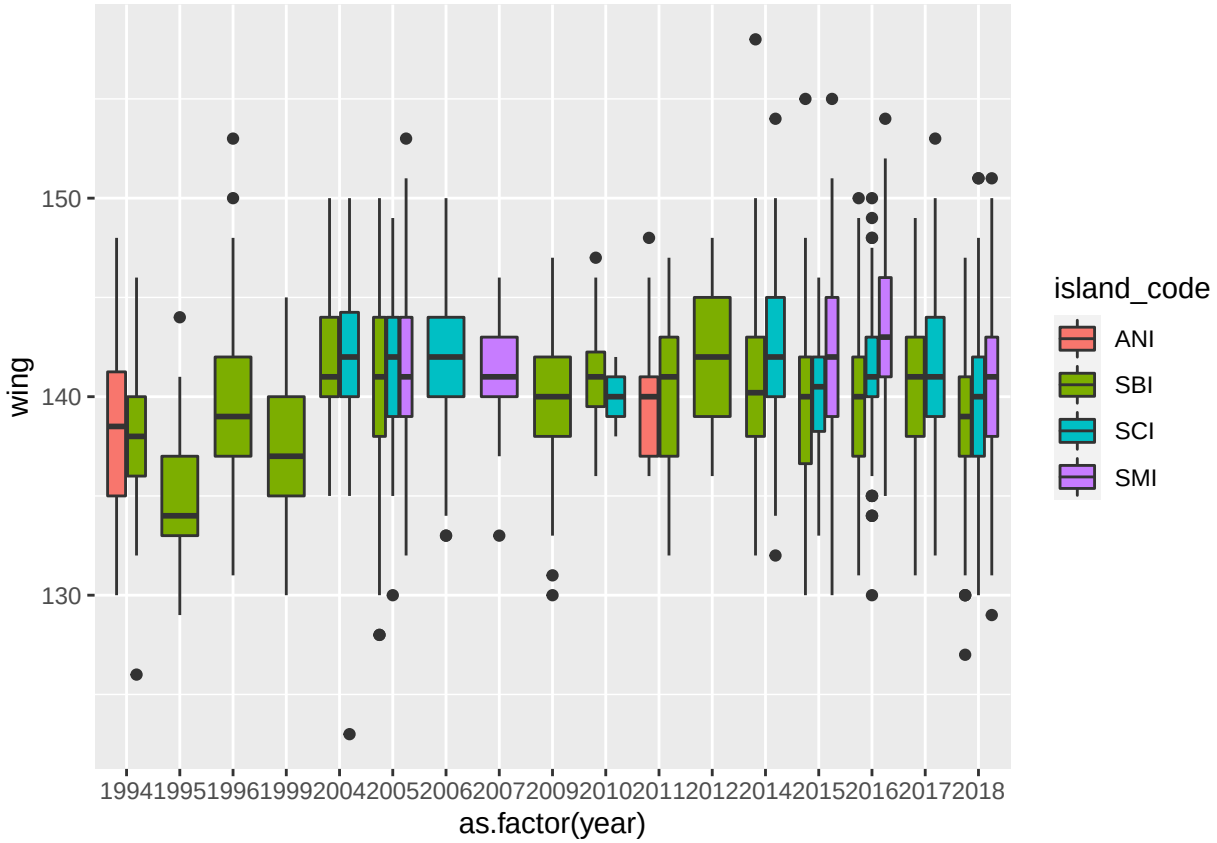



```
# Values in earlier years (~1990s) appear smaller.

# How about mean wing chord observed per year?
WC_yr <- group_by(.data = ASSP, year) %>%
          summarise(wing_av = mean(wing, na.rm = TRUE)) %>%
          ungroup()

ggplot(WC_yr) +
  geom_point(aes(x = year, y = wing_av))
```




```
# Boxplot of wing chord values
ggplot(ASSP, aes(x = as.factor(year), y = wing, fill = island_code)) +
        geom_boxplot()
```

```
## Warning: Removed 161 rows containing non-finite values (stat_boxplot).
```



It does look like the wing chord values recorded in the 1990s on Santa Barbara Island (PRBO) and Anacapa Island (Harry Carter) are significantly smaller than the values recorded by USGS and CHIS collaborators in the 2000-2010s. It's possible they used a relaxed wing measurement as opposed to a flattened wing measurement used later on. We made a note in the Banding Data Dictionary under "wing" that a different method might have been used to take wing chord measurements over time.

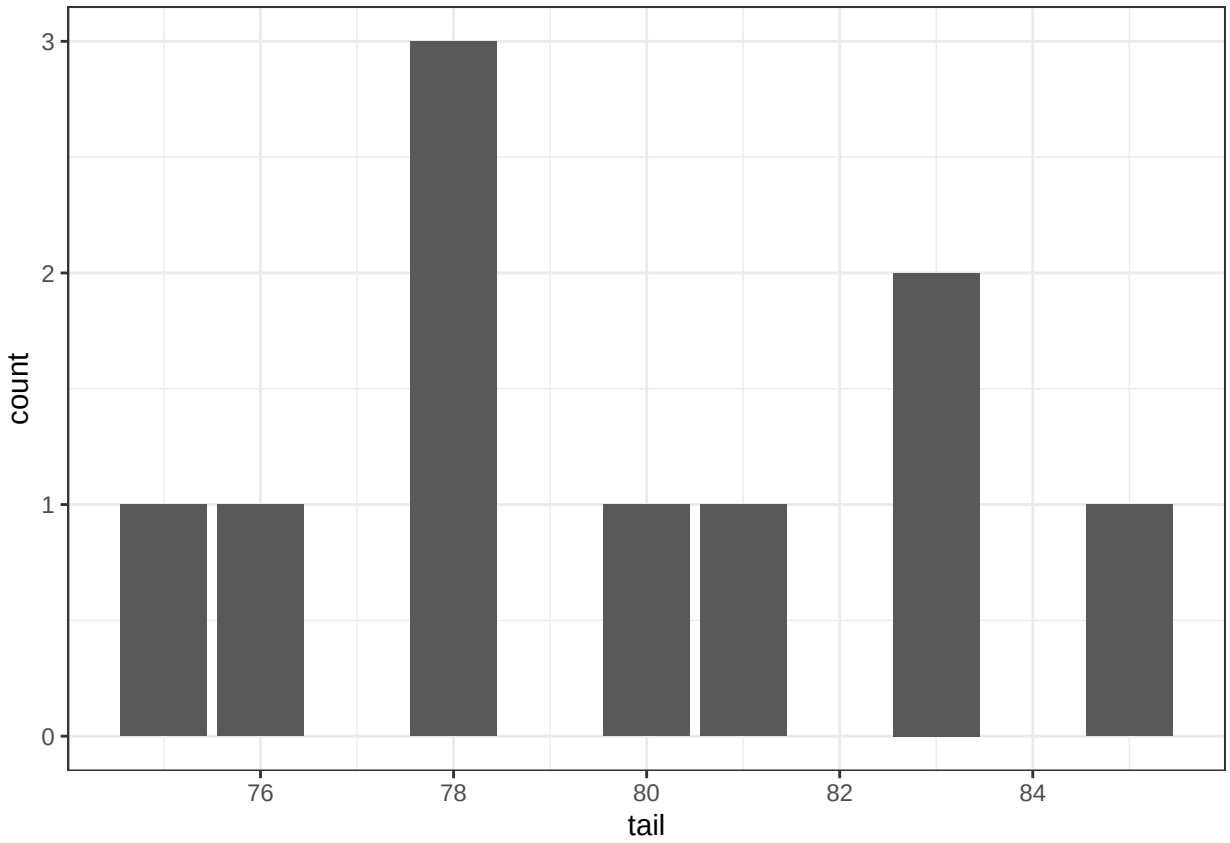
Tail

```
summary(ASSP$tail)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    75.0    78.0    79.0    79.7    82.5    85.0    3833
```

```
# Large range in values recorded.

ggplot(data = ASSP) +
  geom_bar(mapping = aes(x = tail)) +
  theme_bw()
```

```
## Warning: Removed 3833 rows containing non-finite values (stat_count).
```




```
# Values reported in Pyle guide: 72-84 (95% CI).

#Isolate questionable data.
tail_chk <- filter(ASSP, tail < 71 | tail > 85)

# Export to csv and cross-reference raw data.
#write.csv(tail_chk, "tailQAQC.csv")
```

We reviewed the exported data and cross-reference it with raw data (if available). If there was a data entry error, we rectified the mistake in the database and tracked changes in a separate csv file.

CPUE

Set-up data for analysis.

```
metadata <- cpue %>%
              mutate_at(c("app_sunset", "std_ending"),
                        .funs = ~as.POSIXct(., format="%m/%d/%Y %H:%M")) %>%
              mutate_at(c("net_open_1", "net_close_1", "net_open_2", "net_close_2", "net_open_3",
                          "net_close_3", "net_open_4", "net_close_4", "net_open_5", "net_close_5"),
                        .funs = ~as.POSIXct(., format="%Y-%m-%d %H:%M:%S")) %>%
              mutate_at(c("app_sunset", "std_ending", "net_open_1", "net_close_1"),
```

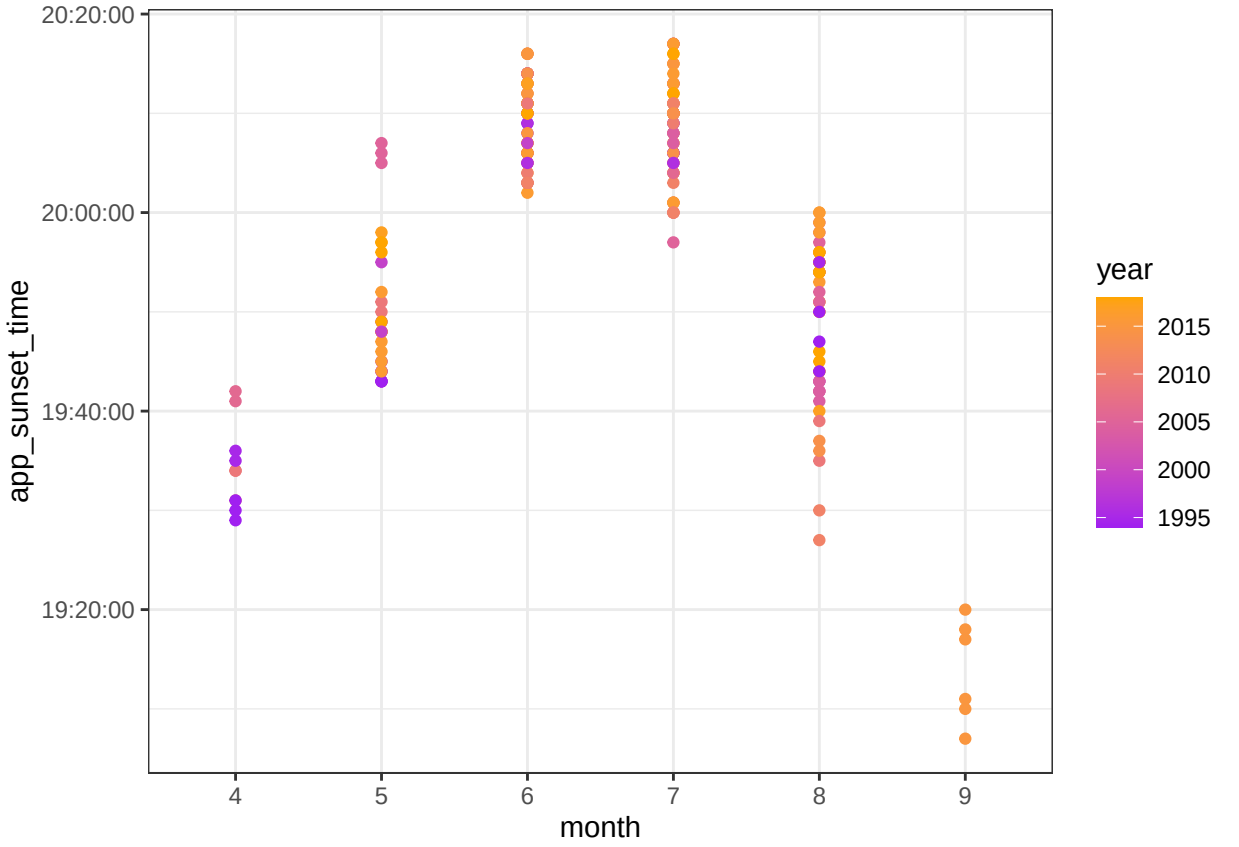
```
                        .funs = list(time = ~ hms::as_hms(.))) %>%
              mutate(CPUE_ratio = CPUEstd/CPUEraw,
                     month = as.character(month)) %>%
              filter(TRUE)

metadata_effort <- metadata %>%
                       select(session_ID, app_sunset, std_ending)
```

Time and Mistnetting Effort

Graphical check of App_sunset

```
ggplot(metadata, aes(month, app_sunset_time)) +
  geom_point(aes(color = year)) +
  scale_color_gradient(low="purple", high="orange") +
  theme_bw()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



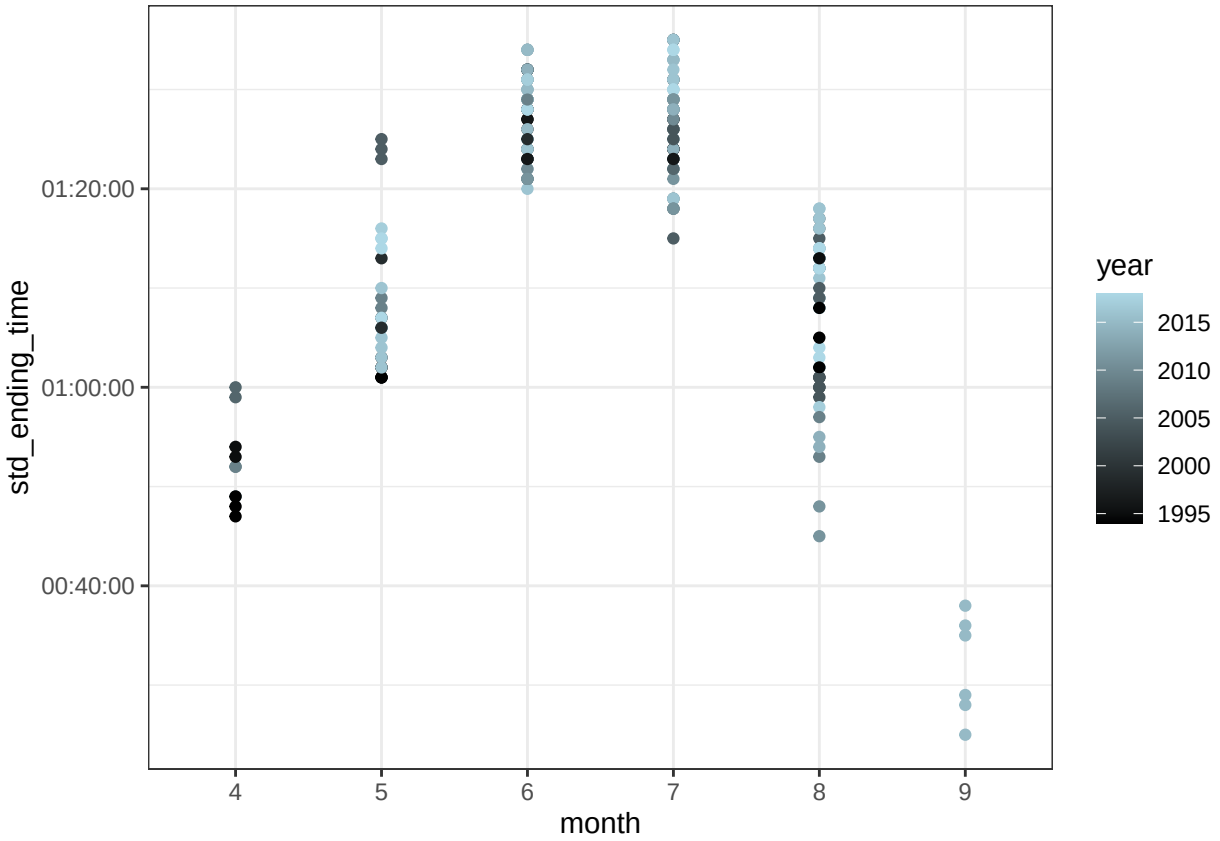
This graph shows the time of apparent sunset for netting sessions each month. The range and timing for that time of year is as we would expect. Thus we conclude that the suncalc function was used effectively to get the sunset times associated with each mistnetting session.

Graphical check of Std_ending

Plotted by month

```
ggplot(metadata, aes(month, std_ending_time)) +
  geom_point(aes(color = year)) +
  scale_color_gradient(low="black", high="light blue") +
  theme_bw()
```

```
## Warning: Removed 2 rows containing missing values (geom_point).
```



This graph shows the time of standard ending (5.3 hours after sunset) for netting sessions each month. The range and timing of the standard ending track with sunset time as we would expect.

Summarize net_open and net_close

```
# first (and usually only) net open time
summary(as.POSIXct(metadata$net_open_1_time))
```

```
##                  Min.               1st Qu.                Median
## "1970-01-01 00:00:00" "1970-01-01 20:45:00" "1970-01-01 21:02:30"
##                  Mean               3rd Qu.                  Max.
```

```
## "1970-01-01 20:33:17" "1970-01-01 21:35:45" "1970-01-01 23:35:00"
##                   NA's
##                   "22"
```

```
# first (and usually only) net close time
summary(as.POSIXct(metadata$net_close_1_time))
```

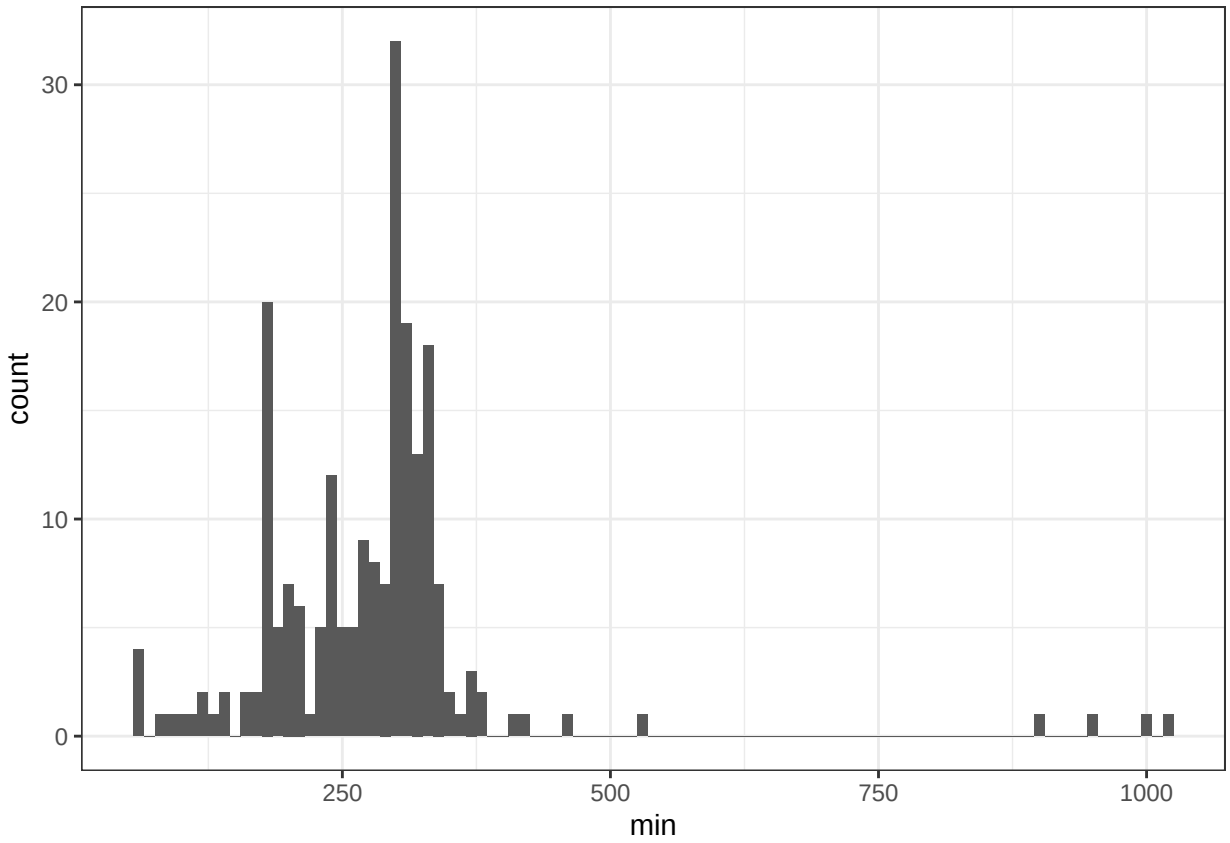
```
##                   Min.                1st Qu.                 Median
## "1970-01-01 00:00:00" "1970-01-01 01:25:00" "1970-01-01 02:00:00"
##                   Mean                3rd Qu.                   Max.
## "1970-01-01 03:43:47" "1970-01-01 02:20:00" "1970-01-01 23:59:00"
##                   NA's
##                   "21"
```

This is not a perfect way to summarize net open and close times because the "summarize" function doesn't recognize times across midnight here. But, by looking at the median and mean, we can tell that net open and close times are usually what we would expect, with a few late/early nights thrown in.

Total mistnetting minutes per session

```
ggplot(metadata, aes(min)) +
  geom_histogram(binwidth = 10) +
  theme_bw()
```

```
## Warning: Removed 24 rows containing non-finite values (stat_bin).
```




```
# summary of total mistnetting minutes
summary(metadata$min)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    56.0   213.8   293.0   280.3   316.0  1022.0      24
```

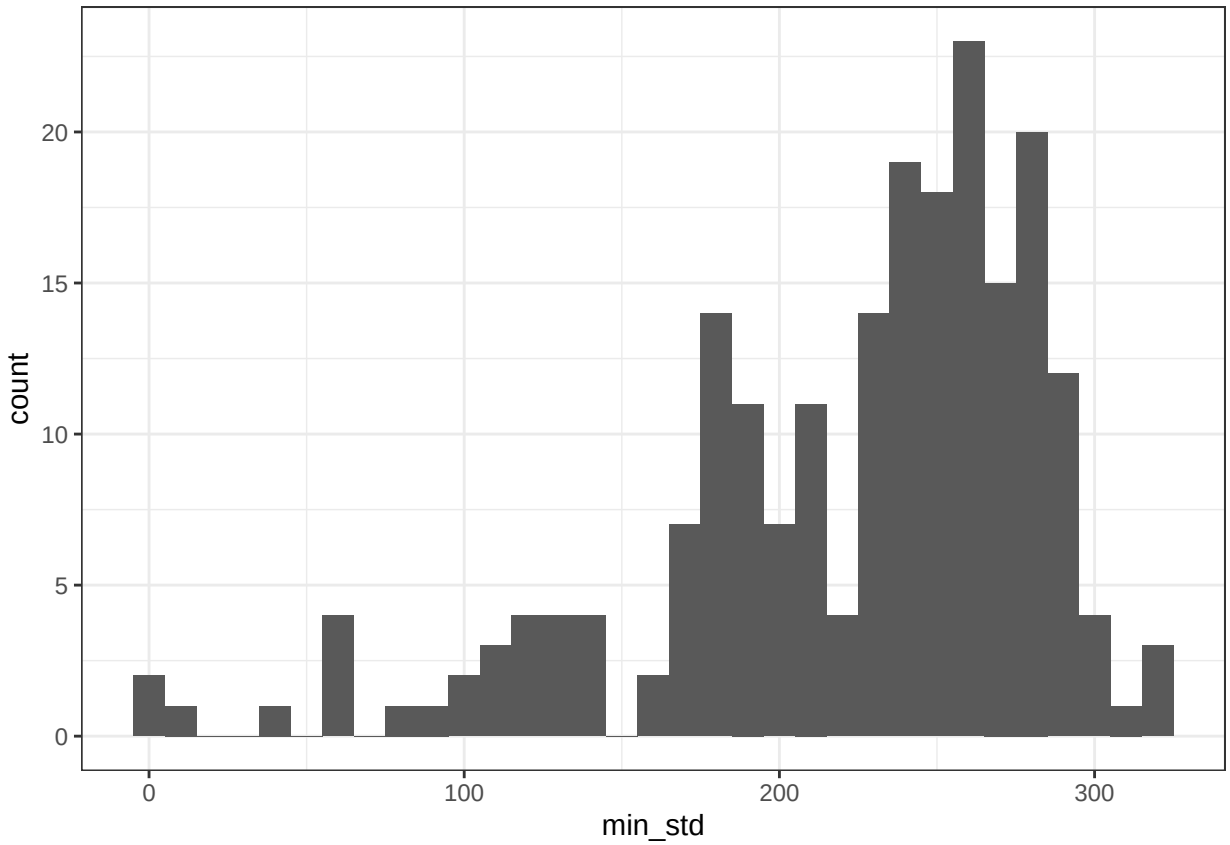
The graph and summary stats above show minutes calcuated for each netting session. We want to check that minutes were added correctly across multiple open/close sessions and also that minutes were added accurately across midnight. It looks like minutes were not added accurately across midnight on four occasions (the outliers to the far right)

Total mistnetting standard minutes per session

from start until end or standard ending, whichever came first

```
ggplot(metadata, aes(min_std)) +
  geom_histogram(binwidth = 10) +
  theme_bw()
```

```
## Warning: Removed 24 rows containing non-finite values (stat_bin).
```




```
# summary of mistnetting minutes cut to standard ending time
summary(metadata$min_std)
```

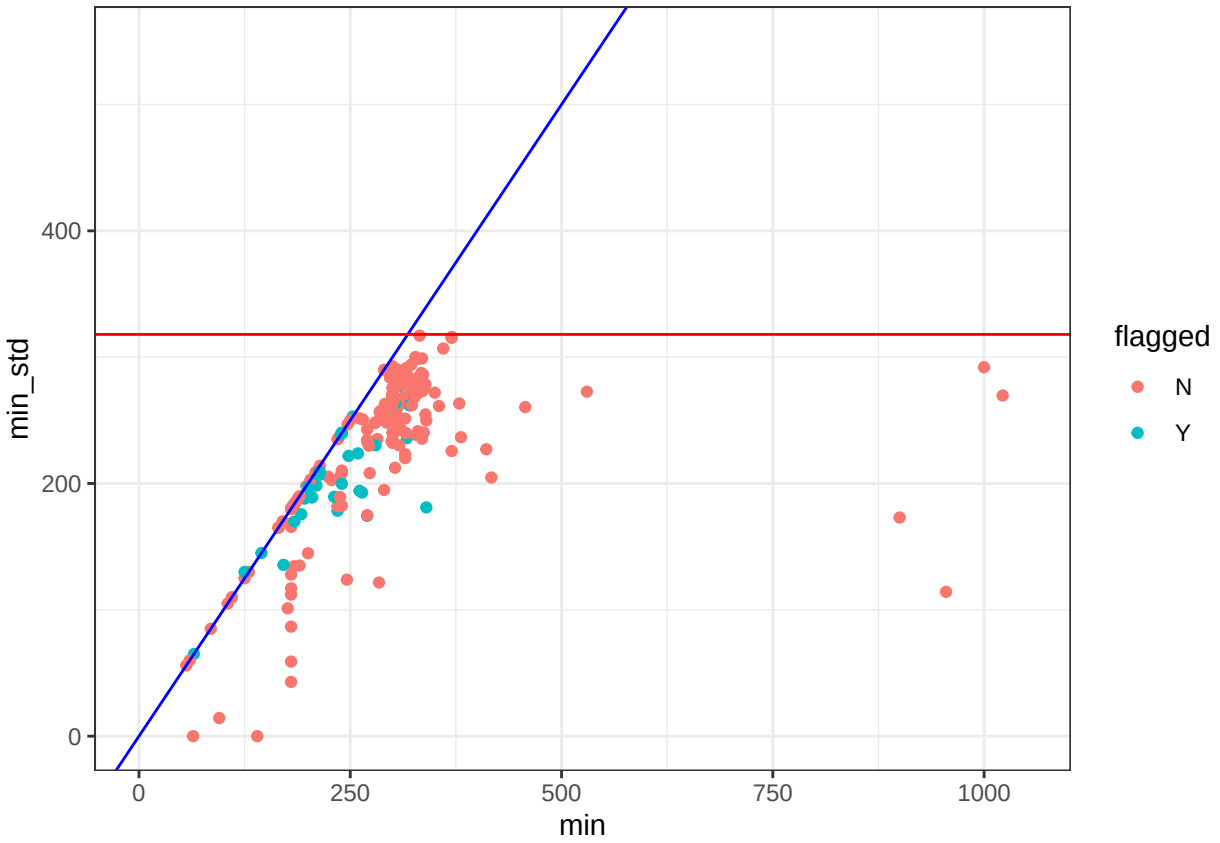
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0   188.8   239.6   222.8   267.7   317.1      24
```

The above graph and summary stats show the total number of mintues for the standardized session (from net open to net close or standard ending whichever came first). The standard ending is 5.3 hours (318 minutes) after sunset. Nets opened sometime after sunset.

Compare min vs. min_std for each session

```
ggplot(metadata, aes(min, min_std)) +
  geom_point(aes(color = flagged)) +
  geom_abline(intercept = 0, slope = 1, color = "blue") +
  geom_hline(yintercept = 318, color = "red") +
  xlim(0,1050) + ylim(0, 550) +
  theme_bw()
```

```
## Warning: Removed 24 rows containing missing values (geom_point).
```



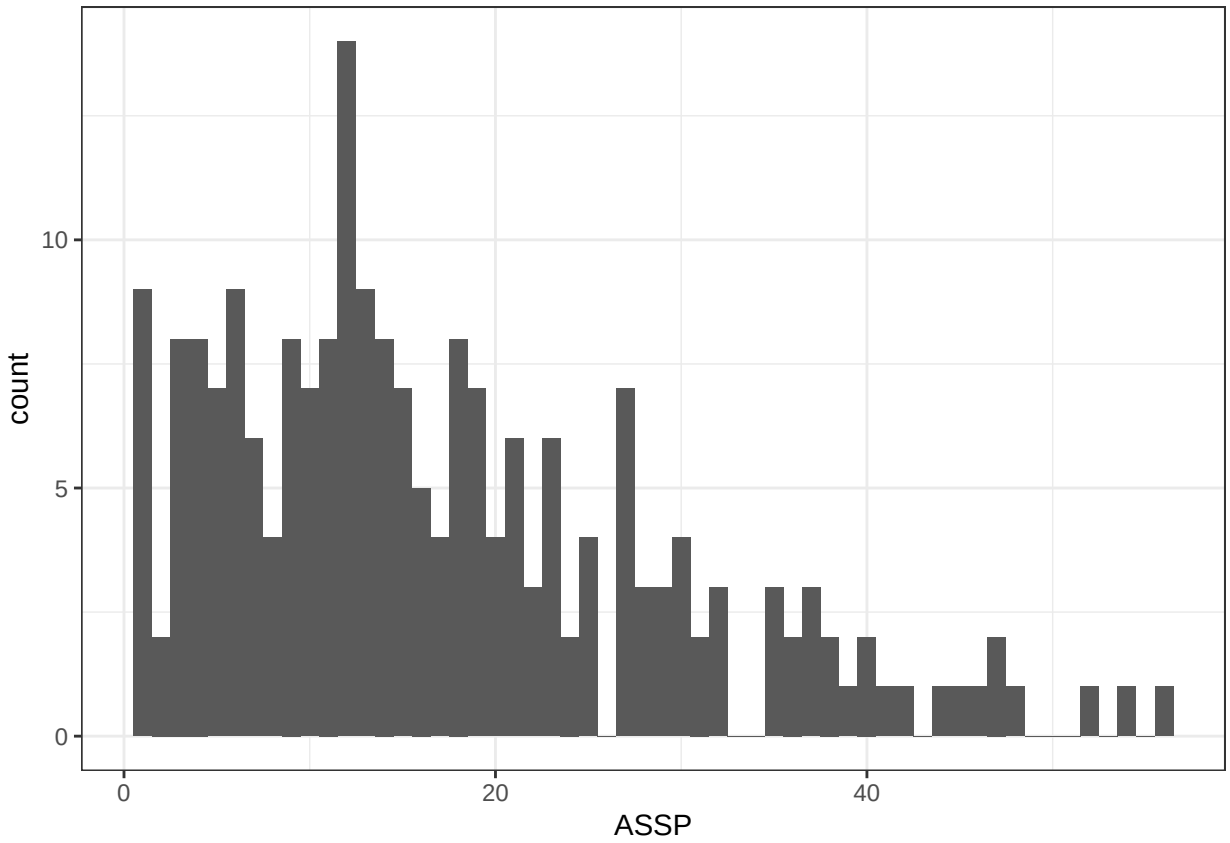
This plot of minutes vs. standardized mintues (before the 5.3 hour standardized ending). Blue line = slope of 1. Red line = standard ending. Here we can make sure standardized net open minutes is equal or less the total number of minutes and sunset to standard ending of 5.3 hours (318 minutes). Red points = data that has been flagged due to inconsistencies in data entry. It does not appear that the reason these entries were flagged effects net minutes.

ASSP

Histogram of total ASSP caught per session

```
ggplot(metadata, aes(ASSP)) +
  geom_histogram(binwidth = 1) +
  theme_bw()
```

```
## Warning: Removed 27 rows containing non-finite values (stat_bin).
```




```
# summary of ASSP catches
summary(metadata$ASSP)
```

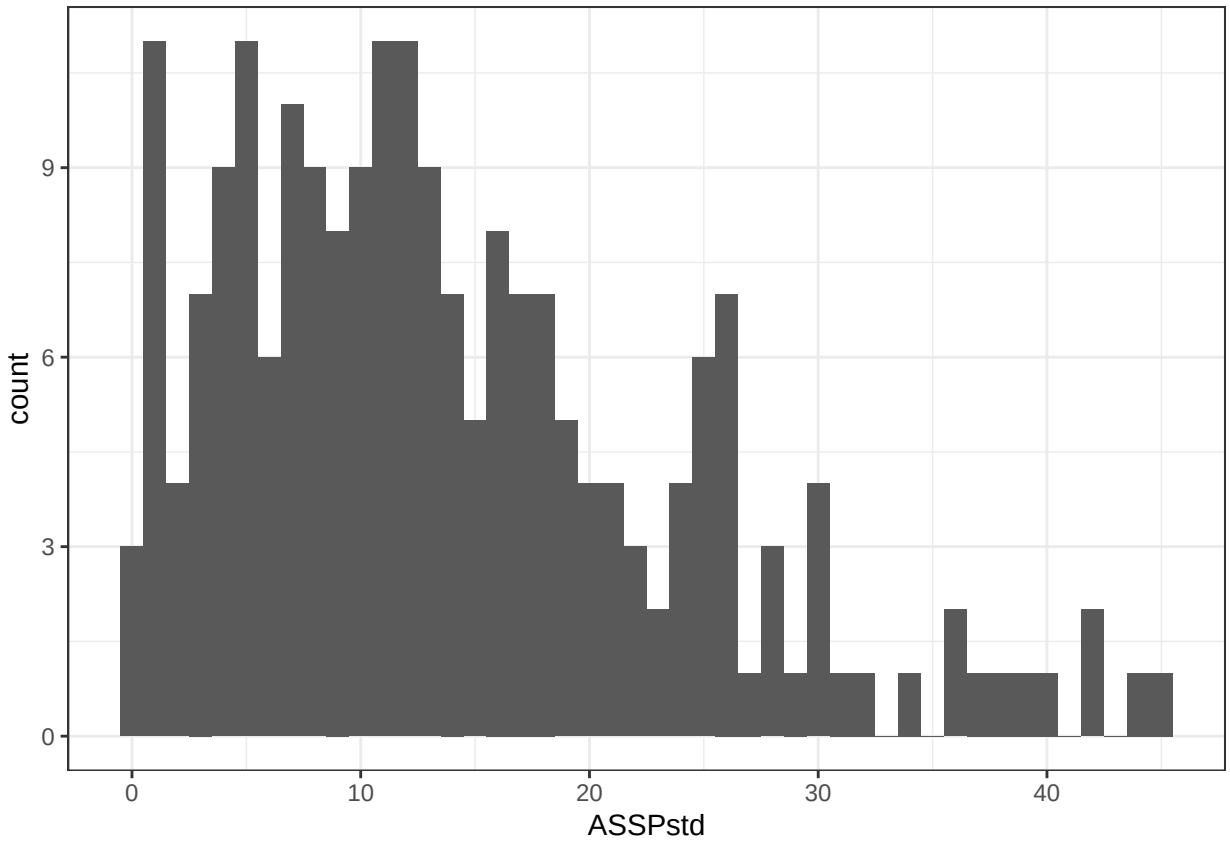
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.00    8.00   14.00   17.18   23.00   56.00      27
```

The graph and summary stats above show the distribution of total numbers of ASSP caught per session.

Histogram of total ASSP caught per standardized session

```
ggplot(metadata, aes(ASSPstd)) +
  geom_histogram(binwidth = 1) +
  theme_bw()
```

```
## Warning: Removed 27 rows containing non-finite values (stat_bin).
```




```
# summary of standardized ASSP catches
summary(metadata$ASSPstd)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    7.00   12.00   14.08   19.00   45.00      27
```

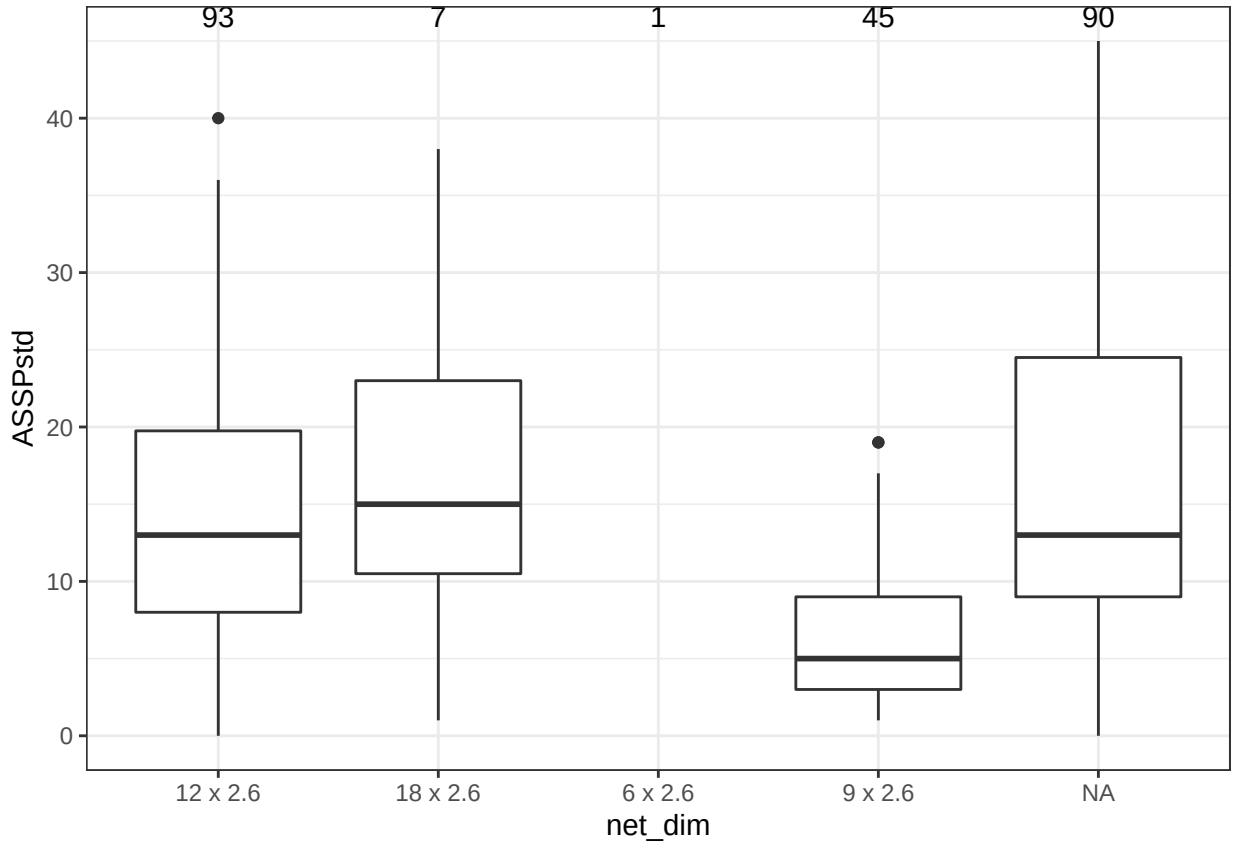
The graph and summary stats above show the distribution of total numbers of ASSP caught before standard ending or net close, whichever came first. This distribution is more constrained than the one above, which is what we would expect with the standard ending cutoff.

Next we will look into what else could effect the number of birds caught.

Number of ASSP caught in relation to net size

```
netUse <- metadata %>%
  group_by(net_dim) %>%
  tally()

ggplot(metadata, aes(net_dim, ASSPstd)) +
  geom_boxplot() +
  geom_text(data = netUse,
            aes(net_dim, Inf, label = n), vjust = 1) +
  theme_bw()
```

```
## Warning: Removed 27 rows containing non-finite values (stat_boxplot).
```

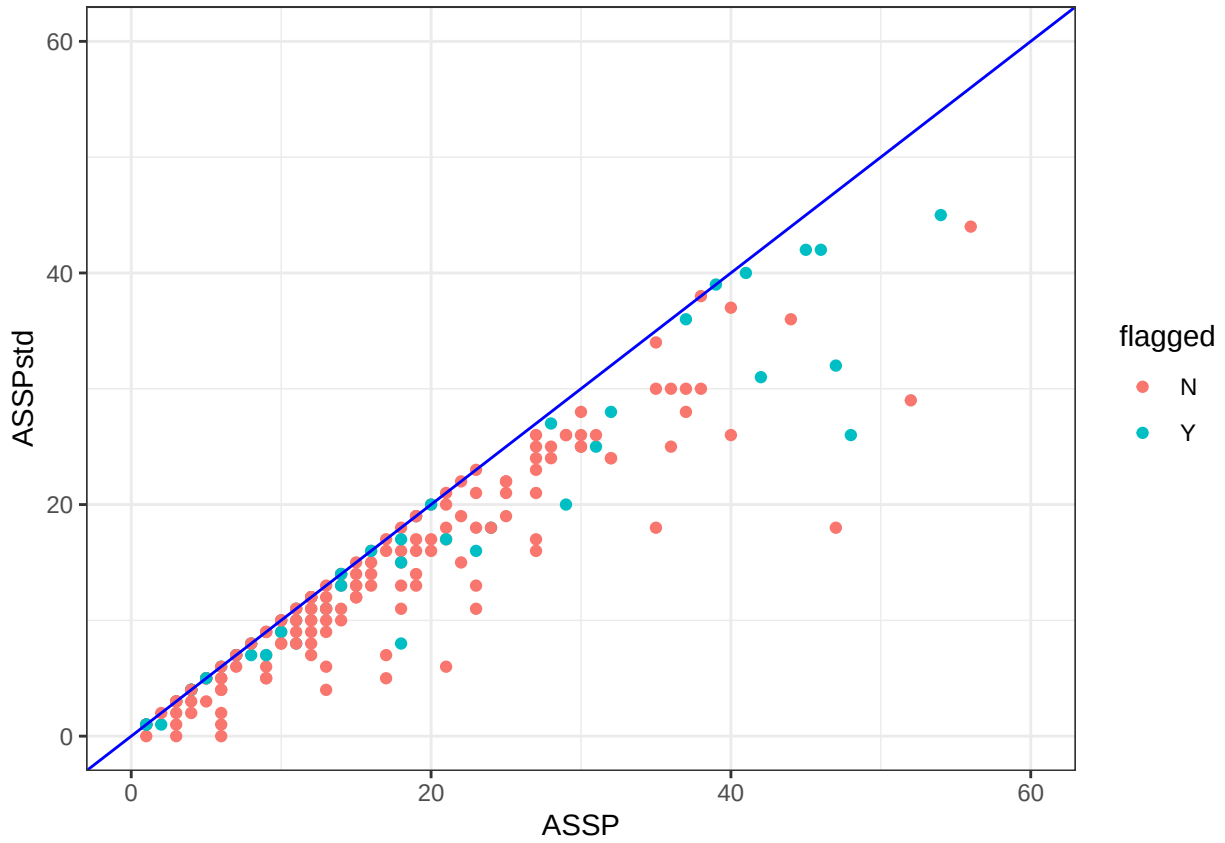


The graph above shows the number of ASSP caught within the standardized period in relation to net dimensions. Number at top of each box plot = sample size. The size of the net doesn't seem to effect the number of birds caught in a specific night. Below we will explore this effect with number of catches standardized to effort.

Comparison of ASSP vs ASSPstd

```
ggplot(metadata, aes(ASSP, ASSPstd)) +
  geom_point(aes(color = flagged)) +
  geom_abline(intercept = 0, slope = 1, color = "blue") +
  xlim(0,60) + ylim(0, 60) +
  theme_bw()
```

```
## Warning: Removed 27 rows containing missing values (geom_point).
```



The plot above shows the total number of ASSP vs. total number of ASSP before the standard ending. Blue line = slope of 1. Here we double check that the standardized number of ASSP is always equal to or less than the total number. Red points = data that has been flagged due to inconsistencies in data entry. It does not appear that the reason these entries were flagged effects the number of birds caught.

Timing of ASSP catches

Next lets explore the frequency of catches in relation to the standard ending. Do catches start dropping off before 5.3 hours after sunset? After? NOTE - the timing of net closures will effect the number of late night captures

```
{r} # ggplot(catches, aes(catchPastSS)) + #   geom_histogram(binwidth
= 10) + #   geom_vline(xintercept = 318, color = "red") + #
xlab("Time past Sunset (min)") + ylab("Number of ASSP Catches")
+ #   theme_bw() #
```

The 5.3 hour cutoff occurs when number of catches are still relatively high. How does that differ across years? In recent years (2014 - 2018) nets were usually closed at 2am. In earlier years nets were sometimes left open past 2am.

```
{r} # ggplot(catches, aes(catchPastSS)) + #   geom_histogram(binwidth
= 10) + #   geom_vline(xintercept = 318, color = "red") + #
xlab("Time past Sunset (min)") + ylab("Number of ASSP Catches")
+ #   facet_wrap(.~ year, scales = "free") + #   theme_bw() #
# NOTE: scales on x- and y- axis differ between panes
```