Data analysis project

Jess Qualley and Rebecca Hansen

March 27, 2020

##Clear code (**delete before submitting**\*)

rm(list = ls(all = TRUE))

#Preparation ##Loading and citing packages

#Load faraway package  
library(faraway)  
  
#Load plyr package  
library(plyr)

##   
## Attaching package: 'plyr'

## The following object is masked from 'package:faraway':  
##   
## ozone

#Load tidyverse package  
library(tidyverse)

## ── Attaching packages ────────────────────────────────────────────────────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✔ ggplot2 3.2.1 ✔ purrr 0.3.3  
## ✔ tibble 2.1.3 ✔ dplyr 0.8.3  
## ✔ tidyr 1.0.2 ✔ stringr 1.4.0  
## ✔ readr 1.3.1 ✔ forcats 0.4.0

## ── Conflicts ───────────────────────────────────────────────────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::arrange() masks plyr::arrange()  
## ✖ purrr::compact() masks plyr::compact()  
## ✖ dplyr::count() masks plyr::count()  
## ✖ dplyr::failwith() masks plyr::failwith()  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::id() masks plyr::id()  
## ✖ dplyr::lag() masks stats::lag()  
## ✖ dplyr::mutate() masks plyr::mutate()  
## ✖ dplyr::rename() masks plyr::rename()  
## ✖ dplyr::summarise() masks plyr::summarise()  
## ✖ dplyr::summarize() masks plyr::summarize()

#Cite packages  
citation("faraway")

##   
## To cite package 'faraway' in publications use:  
##   
## Julian Faraway (2016). faraway: Functions and Datasets for Books  
## by Julian Faraway. R package version 1.0.7.  
## https://CRAN.R-project.org/package=faraway  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Manual{,  
## title = {faraway: Functions and Datasets for Books by Julian Faraway},  
## author = {Julian Faraway},  
## year = {2016},  
## note = {R package version 1.0.7},  
## url = {https://CRAN.R-project.org/package=faraway},  
## }  
##   
## ATTENTION: This citation information has been auto-generated from  
## the package DESCRIPTION file and may need manual editing, see  
## 'help("citation")'.

citation("plyr")

##   
## To cite plyr in publications use:  
##   
## Hadley Wickham (2011). The Split-Apply-Combine Strategy for Data  
## Analysis. Journal of Statistical Software, 40(1), 1-29. URL  
## http://www.jstatsoft.org/v40/i01/.  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {The Split-Apply-Combine Strategy for Data Analysis},  
## author = {Hadley Wickham},  
## journal = {Journal of Statistical Software},  
## year = {2011},  
## volume = {40},  
## number = {1},  
## pages = {1--29},  
## url = {http://www.jstatsoft.org/v40/i01/},  
## }

citation("tidyverse")

##   
## Wickham et al., (2019). Welcome to the tidyverse. Journal of  
## Open Source Software, 4(43), 1686,  
## https://doi.org/10.21105/joss.01686  
##   
## A BibTeX entry for LaTeX users is  
##   
## @Article{,  
## title = {Welcome to the {tidyverse}},  
## author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostino McGowan and Romain François and Garrett Grolemund and Alex Hayes and Lionel Henry and Jim Hester and Max Kuhn and Thomas Lin Pedersen and Evan Miller and Stephan Milton Bache and Kirill Müller and Jeroen Ooms and David Robinson and Dana Paige Seidel and Vitalie Spinu and Kohske Takahashi and Davis Vaughan and Claus Wilke and Kara Woo and Hiroaki Yutani},  
## year = {2019},  
## journal = {Journal of Open Source Software},  
## volume = {4},  
## number = {43},  
## pages = {1686},  
## doi = {10.21105/joss.01686},  
## }

##Read in and organize data

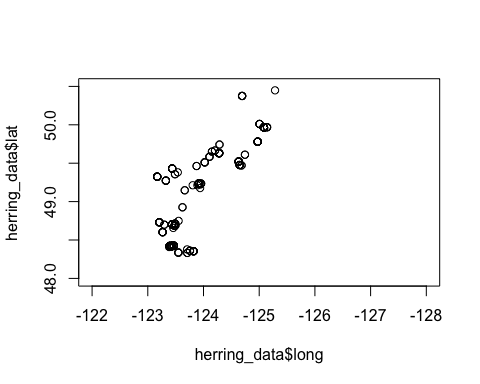
#Read from csv file  
herring <- read.csv("HerringOtolithDatabase 15 Jan 2020\_JQ.csv", header=TRUE, fileEncoding="UTF-8-BOM")  
  
#Create a data frame  
herring\_data <- data.frame(sal.id = herring$FishCode,  
 oto.width = herring$AverageWidth,  
 digestion = herring$Digestion,  
 sal.sp = herring$SalmonSpecies,  
 sal.sex = herring$SalmonSex,  
 sal.length = as.numeric(herring$SalmonLength),  
 sal.weight = herring$SalmonWeight,  
 coll.year = herring$CollectionYear,  
 coll.doy = herring$CollectionDayofYear,  
 lat = herring$Latitude,  
 long = herring$Longitude,  
 stat.area = as.factor(herring$StatArea))  
  
#figure out which stat areas to remove  
levels(as.factor(herring$StatArea))

## [1] "1" "13" "14" "15" "16" "17" "18" "19" "20" "23" "25"   
## [12] "28" "29" "101" "125"

#filter for 2018 data, chinook salmon only, omit northern BC areas (Haid Gwaii) and west coast Vancouver Island  
herring\_data <- herring\_data %>% filter(herring\_data$sal.sp == "ch" & herring\_data$coll.year == "2018" & herring\_data$stat.area != "1" & herring\_data$stat.area != "101" & herring\_data$stat.area != "125" & herring\_data$stat.area != "23" & herring\_data$stat.area != "25") %>% na.omit()

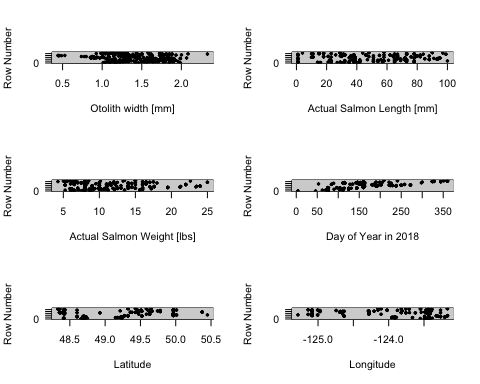
##Checking our work (***delete before submitting***)

#plot lat/long as continuous variable  
plot(herring\_data$long, herring\_data$lat, xlim = c(-122, -128), ylim = c(48, 50.5))



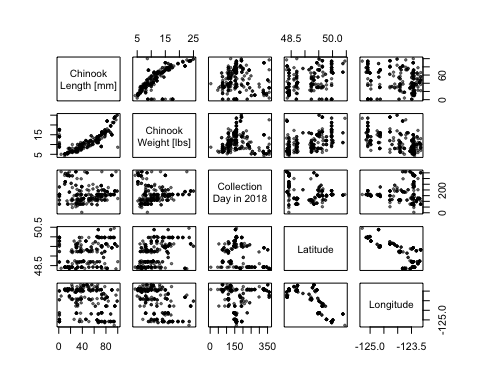
##Example 1 (\*\*\*delete before submitting)

#add column for number of rows  
n <- nrow(herring\_data)  
herring\_data$row <- 1:n  
  
par(mfrow = c(3, 2))  
  
#otolith width  
with(herring\_data, plot(oto.width, row, las = 1, type = "n", xlab = "Otolith width [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(oto.width, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#actual salmon length - all salmon that ate herring - note that some numbers are ranges and have weird symbols...  
#where lengths were not recorded, they were calculated using length-weight relationship  
with(herring\_data, plot(sal.length, row, las = 1, type = "n", xlab = "Actual Salmon Length [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.length, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#actual salmon weight - raw salmon weights no back calculations  
with(herring\_data, plot(sal.weight, row, las = 1, type = "n", xlab = "Actual Salmon Weight [lbs]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.weight, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#day of year  
with(herring\_data, plot(coll.doy, row, las = 1, type = "n", xlab = "Day of Year in 2018", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(coll.doy, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#lat  
with(herring\_data, plot(lat, row, las = 1, type = "n", xlab = "Latitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(lat, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#long  
with(herring\_data, plot(long, row, las = 1, type = "n", xlab = "Longitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(long, row, las = 1, type = "p", pch = 19, cex = 0.5))



##Example 2 (\*\*\*delete before submitting)

#create data frame for pairwise scatter with only continuous explanatory variables - salmon lenght, weight, collection day of year, latitude and longitude  
herring\_collinearity <- data.frame("sal.length" = herring\_data$sal.length, "sal.weight" = herring\_data$sal.weight, "coll.doy" = herring\_data$coll.doy, "lat" = herring\_data$lat, "long" = herring\_data$long)  
  
#create pairwise scatterplot with salmon length, weight and day of year  
plot(herring\_collinearity[1:5], cex = 0.5, pch = 19, col = rgb(0, 0, 0, 0.5),   
 labels = c("Chinook\nLength [mm]", "Chinook\nWeight [lbs]", "Collection\nDay in 2018", "Latitude", "Longitude"))



#pairwise correlation coefficients for multiple variables  
print(cor(na.omit(herring\_data[ ,c("sal.length", "sal.weight", "coll.doy", "lat", "long")])), digits = 2)

## sal.length sal.weight coll.doy lat long  
## sal.length 1.000 0.71 -0.074 0.32 -0.22  
## sal.weight 0.709 1.00 -0.114 0.50 -0.51  
## coll.doy -0.074 -0.11 1.000 -0.28 0.14  
## lat 0.321 0.50 -0.277 1.00 -0.84  
## long -0.225 -0.51 0.137 -0.84 1.00

#variance inflation factors  
print(vif(na.omit(herring\_data[ ,c("sal.length", "sal.weight", "coll.doy", "lat", "long")])))

## sal.length sal.weight coll.doy lat long   
## 2.237120 2.748406 1.126163 4.076674 4.144454

##Example 3 (\*\*\*delete before submitting)

#Make a table of salmon sex versus year  
with(herring\_data, table(sal.sex, coll.year))

## coll.year  
## sal.sex 2018  
## 46  
## female 137  
## male 100

#Make a table of salmon sex versus digestion  
with(herring\_data, table(sal.sex, digestion))

## digestion  
## sal.sex 1 2 3 4  
## 2 4 19 21  
## female 2 25 62 48  
## male 2 17 52 29

#Make a table of salmon year versus digestion  
with(herring\_data, table(coll.year, digestion))

## digestion  
## coll.year 1 2 3 4  
## 2018 6 46 133 98

##Example 4 (\*\*\*delete before submitting)

#Make a table of 0 values and NAs for oto.width  
table(herring\_data$oto.width > 0, useNA = "always")

##   
## TRUE <NA>   
## 283 0

#Make a table of 0 values and NAs for sal.length  
table(herring\_data$sal.length > 0, useNA = "always")

##   
## TRUE <NA>   
## 283 0

#Make a table of 0 values and NAs for sal.weight  
table(herring\_data$sal.weight > 0, useNA = "always")

##   
## TRUE <NA>   
## 283 0

#Make a table of 0 values and NAs for lat  
table(herring\_data$lat > 0, useNA = "always")

##   
## TRUE <NA>   
## 283 0

#Make a table of 0 values and NAs for long  
table(herring\_data$long > 0, useNA = "always")

##   
## FALSE <NA>   
## 283 0

#Proportion of measured oto-width values that are zero:  
sum(herring\_data$oto.width == 0, na.rm = TRUE) / length(na.omit(herring\_data$oto.width))

## [1] 0

#Proportion of measured sal.length values that are zero:  
sum(herring\_data$sal.length == 0, na.rm = TRUE) / length(na.omit(herring\_data$sal.length))

## [1] 0

#Proportion of measured sal.weight values that are zero:  
sum(herring\_data$sal.weight == 0, na.rm = TRUE) / length(na.omit(herring\_data$sal.weight))

## [1] 0

#Proportion of measured lat values that are zero:  
sum(herring\_data$lat == 0, na.rm = TRUE) / length(na.omit(herring\_data$lat))

## [1] 0

#Proportion of measured long values that are zero:  
sum(herring\_data$long == 0, na.rm = TRUE) / length(na.omit(herring\_data$long))

## [1] 0

##Cleaning data

#filter only chinook and data from 2018 with na removed  
herring\_data\_2018 <- herring\_data %>% filter(herring\_data$sal.sp == "ch" & herring\_data$coll.year == "2018") %>% na.omit()

#Section 1: Introduction, Question, Goals and Hypotheses (400 words) ##A) Introduce relevant concepts: - State of salmon in BC - Salmon diet - Herring as a prey item for salmon - Gape size theory? Gape size and body size? Prey selection strategies?

How the dataset was collected: - Salmon stomachs donated to Juanes lab by anglers - Salmon stomach contents sorted and classfied based on digestion - Otoliths collected and measured

Background needed to understand the dataset: - Otolith width as the best proxy for herring length

At least two scientiﬁc papers that are related to your topic:

##B) Research question: - Relationship between salmon size and size of their herring prey

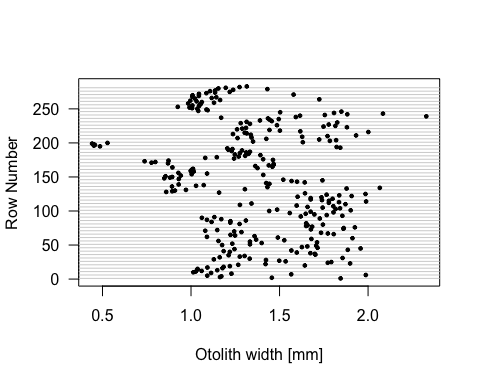
##C) Goal of research: - Inference, we want to test if the size of the salmon determines the size of its herring prey,

##D) Hypothesis: - hypothesis for how you expect the response variable to change with each predictor variable and interactions: linear or non-linear and what shape? positive or negative? interaction terms? Why? OR list of alternative hypotheses, in words or as model formulae that correspond to diﬀerent scenarios related to your question - any null hypotheses

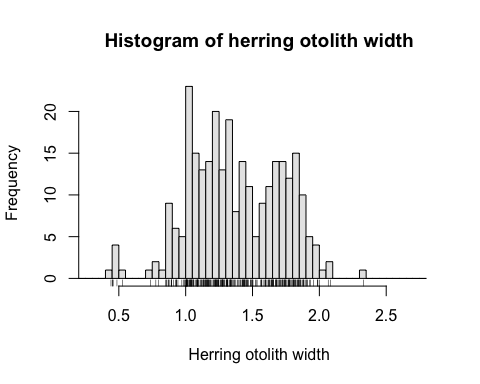
##E) Model selection: - Why or why not use it? - What type to use?

#Section 2: The Response Variable (150 words) ##A)

#Add column for number of rows  
n <- nrow(herring\_data)  
herring\_data$row <- 1:n  
  
#Dotchart of otolith width (response variable)  
with(herring\_data, plot(oto.width, row, las = 1, type = "n", xlab = "Otolith width [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(oto.width, row, las = 1, type = "p", pch = 19, cex = 0.5))



#Histogram of response variable with rugmarks  
hist(herring\_data$oto.width, xlim = c(0.3, 2.8), breaks = seq(0.2, 2.8, by = 0.05), col = grey(0.9), main = "Histogram of herring otolith width", xlab = "Herring otolith width")  
rug(herring\_data$oto.width)



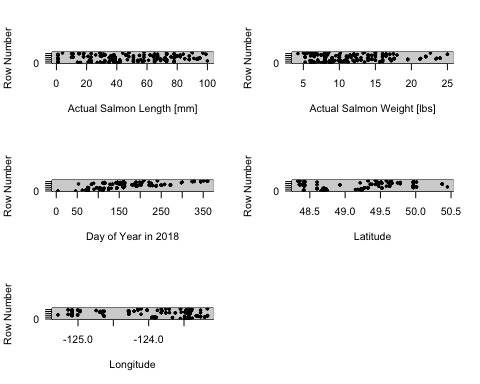
##B) Qualitative description of response variable: - Continuous - Bounded by 0, can’t take negative values

##C) Potential issues with response variable - Outliers - Missing values - Zeroes - Non-independence from repeated measurements within experimental units - Non-idependence from temporal autocorrelation - Non-independence from spatial autocorrelation

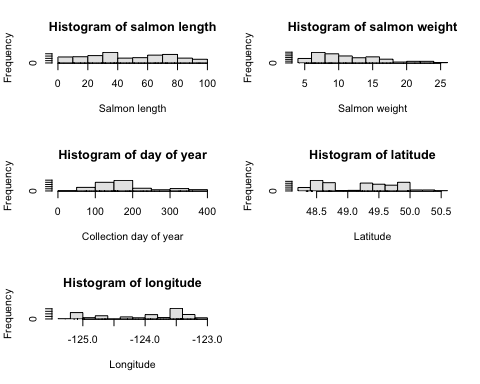
##D) Candidate distributions for variable - Family of probability distributions we’re likely to use

#Section 3: The Explanatory Variables (150 words) ##A) ###Continuous explanatory variables

#Multipanel of dotcharts for continuous explanatory variables  
par(mfrow = c(3, 2))  
  
#Salmon length  
with(herring\_data, plot(sal.length, row, las = 1, type = "n", xlab = "Actual Salmon Length [mm]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.length, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Salmon weight  
with(herring\_data, plot(sal.weight, row, las = 1, type = "n", xlab = "Actual Salmon Weight [lbs]", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.weight, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Collection day of year  
with(herring\_data, plot(coll.doy, row, las = 1, type = "n", xlab = "Day of Year in 2018", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(coll.doy, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Latitude  
with(herring\_data, plot(lat, row, las = 1, type = "n", xlab = "Latitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(lat, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Longitude  
with(herring\_data, plot(long, row, las = 1, type = "n", xlab = "Longitude", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(long, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Multipanel of histograms for continuous explanatory variables  
par(mfrow = c(3, 2))

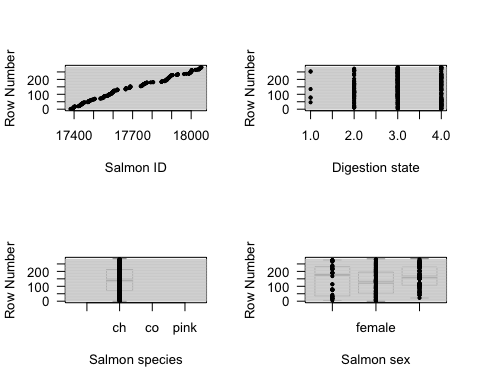


#Salmon length  
hist(herring\_data$sal.length, col = grey(0.9), main = "Histogram of salmon length", xlab = "Salmon length")  
rug(herring\_data$sal.length)  
  
#Salmon weight  
hist(herring\_data$sal.weight, col = grey(0.9), main = "Histogram of salmon weight", xlab = "Salmon weight")  
rug(herring\_data$sal.weight)  
  
#Collection day of year  
hist(herring\_data$coll.doy, col = grey(0.9), main = "Histogram of day of year", xlab = "Collection day of year")  
rug(herring\_data$coll.doy)  
  
#Latitude  
hist(herring\_data$lat, col = grey(0.9), main = "Histogram of latitude", xlab = "Latitude")  
rug(herring\_data$lat)  
  
#Longitude  
hist(herring\_data$long, col = grey(0.9), main = "Histogram of longitude", xlab = "Longitude")  
rug(herring\_data$long)



###Categorical explanatory variables

#Multipanel of dotcharts for categorical explanatory variables  
par(mfrow = c(2, 2))  
  
#Salmon id  
with(herring\_data, plot(sal.id, row, las = 1, type = "n", xlab = "Salmon ID", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.id, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Digestion  
with(herring\_data, plot(digestion, row, las = 1, type = "n", xlab = "Digestion state", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(digestion, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Salmon species  
with(herring\_data, plot(sal.sp, row, las = 1, type = "n", xlab = "Salmon species", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.sp, row, las = 1, type = "p", pch = 19, cex = 0.5))  
  
#Salmon sex  
with(herring\_data, plot(sal.sex, row, las = 1, type = "n", xlab = "Salmon sex", ylab = "Row Number"))  
abline (h = seq(1, n, by = 5), col = "lightgrey")  
with(herring\_data, points(sal.sex, row, las = 1, type = "p", pch = 19, cex = 0.5))



#Tables for categorical explanatory variables  
#Make a table of salmon sex versus year  
with(herring\_data, table(sal.sex, coll.year))

## coll.year  
## sal.sex 2018  
## 46  
## female 137  
## male 100

#Make a table of salmon sex versus digestion  
with(herring\_data, table(sal.sex, digestion))

## digestion  
## sal.sex 1 2 3 4  
## 2 4 19 21  
## female 2 25 62 48  
## male 2 17 52 29

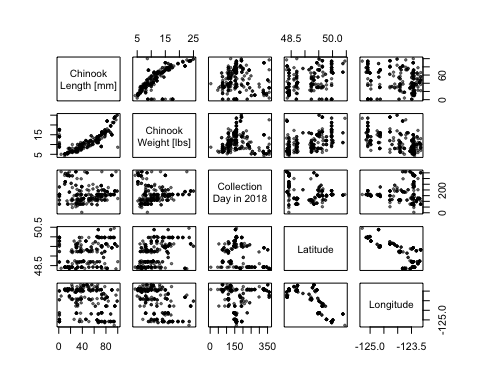
#Make a table of salmon year versus digestion  
with(herring\_data, table(coll.year, digestion))

## digestion  
## coll.year 1 2 3 4  
## 2018 6 46 133 98

##B) - Outliers - Transformations why or why not?

#Section 4: Collinearity, Balance, and Variance Inﬂation Factors (200 words) ##A)

#create data frame for pairwise scatter with only continuous explanatory variables - salmon lenght, weight, collection day of year, latitude and longitude  
herring\_collinearity <- data.frame("sal.length" = herring\_data$sal.length, "sal.weight" = herring\_data$sal.weight, "coll.doy" = herring\_data$coll.doy, "lat" = herring\_data$lat, "long" = herring\_data$long)  
  
#Pairwise scatterplot of explanatory variables (salmon length, weight, day of year, latitude and longitude)  
plot(herring\_collinearity[1:5], cex = 0.5, pch = 19, col = rgb(0, 0, 0, 0.5),   
 labels = c("Chinook\nLength [mm]", "Chinook\nWeight [lbs]", "Collection\nDay in 2018", "Latitude", "Longitude"))



#Pairwise correlation coefficients for explanatory variables  
print(cor(na.omit(herring\_data[ ,c("sal.length", "sal.weight", "coll.doy", "lat", "long")])), digits = 2)

## sal.length sal.weight coll.doy lat long  
## sal.length 1.000 0.71 -0.074 0.32 -0.22  
## sal.weight 0.709 1.00 -0.114 0.50 -0.51  
## coll.doy -0.074 -0.11 1.000 -0.28 0.14  
## lat 0.321 0.50 -0.277 1.00 -0.84  
## long -0.225 -0.51 0.137 -0.84 1.00

#Variance Inflation Factors  
print(vif(na.omit(herring\_data[ ,c("sal.length", "sal.weight", "coll.doy", "lat", "long")])))

## sal.length sal.weight coll.doy lat long   
## 2.237120 2.748406 1.126163 4.076674 4.144454

#Balance of dataset (if appropriate)

Comment on pairwise scatterplots Comment on strength of correlations Comment on VIFs Comment on potential difficulties from correlations and benefits

#Section 5: Statistical Methods and Model Fitting (200 words) Statistical methods - Which you will use and why - Cite packages used for modelling - What family of probability distributions (what link function) - Which explantory variables are we including - If and why data points are excluded - If transformations are used and why - If random-effects are used and why

#Fit the model or models

#Section 6: Model Checking

#Check the assumptions of at least one model  
#- plotting and interpreting the residuals of the model  
#- check for correlation in the residuals due to grouping variables

Summary of the plausibility of the model - Potential outliers

#Section 7: Model Summary, Conﬁdence Intervals and Model Comparison

#Model comparison and/or model selection and/or hypothesis testing  
#- AIC (or AICc) values  
#- AIC (or AICc) weights  
#- R-squared (or adjusted R-squared) values  
#- p-values.   
#Summary for at least one ﬁtted model  
#- coeﬃcient estimates and their standard errors  
#- conﬁdence intervals for coeﬃcient estimate

#Section 8: Plotting a Model with the Data

#Plot at least one model with data  
#- use method that shows strengths and weaknesses  
#- include measurement of uncertainty

#Section 9: Discussion (400 words) ##A) Assessment of how well the model seems to ﬁt the data - model checking - visualization of the model with the data - ##B) Qualitative interpretation of the model - interpretation of results in relation to your research questions and hypotheses - general interpretation of conﬁdence intervals or standard errors for coeﬃcient estimates - ##C) Limitations of the model(s) that you have ﬁt - Limitations - Possible options for improvement - Possible approaches for further data analysis

##D) Results into context - citie at least two relevant scientiﬁc papers