```
#Pratt Info 640 Fall 2019
#Diedre Brown; dbrow207@pratt.edu
#Predictive Data Analysis Project - Due 22 Oct 2019

#call libraries
library(tidyverse)
library(lubridate)
library(dplyr)
library(ggplot2)
library(broom) #broom helps clean things up and remerge dataframes
library(GGally) #GGally hsleps run multiple pair-wise correlations

####Import Datasets####
#Shrimp populations in Quebrada Prieta (Pools 0, 8, 9, 15) (El Verde)
#Source: Crowl T. 2010. Shrimp populations in Quebrada Prieta (Pools 0,
8, 9, 15) (El Verde). Environmental Data Initiative.
#https://doi.org/10.6073/pasta/f6c8497c780ecf619053dcd020d371f2. Dataset
accessed 10/22/2019.
#Creator: Crowl, Todd
#Creator Publication Date: 2010-11-27
#Creator's Abstract:Freshwater shrimp from the Quebrada Prieta (a
tributary to the Sonadora in the Espiritu Santu drainage, have been
censused 6 times yearly since 1988.
#Atya lanipes, Xiphocaris elongata and Macrobrachium spp. are regularly
captured and comprise the species in this data base.

#NOTES ON DATA:
#On further inspection of the creator's notes, I found that the count
figure represents:
#Total number of freshware species of shrimps captured divided by the
number of traps from the corresponding pool in Quebrada Prieta and then
released.
#Number of traps in each pool can vary but usually are 34, 3, and 2 for
Pools 0, 8, 15 respectively. Record is missing when data is missing.
#
#Though biannual and weekly figures were available, I just want an
overall picture to evaluate for correlation and future study, so I will
only use the biannual data.
#Only weekly data since 1993 on Pool 9 was available and therefore not
included.
#

#biannual data for pools 0, 8, and 15 from 1988-2016
shrimppool_0_bia <- read.csv("Datasets/knb-lter-
luq.54.945757/ShrimpPool-0-biannual-1988-2016.csv")
shrimppool_8_bia <- read.csv("Datasets/knb-lter-
luq.54.945757/ShrimpPool-8-biannual-1988-2016.csv")
shrimppool_15_bia <- read.csv("Datasets/knb-lter-
luq.54.945757/ShrimpPool-15-biannual-1988-2016.csv")
```

```
#view data for each pool and check for NA's
#shrimp pool 0
class(shrimppool_0_bia)
head(shrimppool_0_bia)
str(shrimppool_0_bia)
summary(shrimppool_0_bia)
sum(is.na(shrimppool_0_bia))
#shrimp pool 8
class(shrimppool_8_bia)
head(shrimppool_8_bia)
str(shrimppool_8_bia)
summary(shrimppool_8_bia)
sum(is.na(shrimppool_8_bia))
#shrimp pool 15
class(shrimppool_15_bia)
head(shrimppool_15_bia)
str(shrimppool_15_bia)
summary(shrimppool_15_bia)
sum(is.na(shrimppool_15_bia))


####Join All Biannual Dataframes into 1 Wide Dataset and 1 Long
Dataset####
shrimppool_temp_bia = full_join(shrimppool_0_bia, shrimppool_8_bia,
by=c("YEAR", "Month", "POOL", "ATYACPUE", "XIPHCPUE", "MACCPUE" ),
copy=FALSE)
shrimppool_all_bia = full_join(shrimppool_temp_bia, shrimppool_15_bia,
by=c("YEAR", "Month", "POOL", "ATYACPUE", "XIPHCPUE", "MACCPUE" ),
copy=FALSE)
head(shrimppool_all_bia)
str(shrimppool_all_bia)
glimpse(shrimppool_all_bia)
summary(shrimppool_all_bia)
#clean up dataset
#Since POOL is a location description and not a number, let's change
it's type to factor
shrimppool_all_bia$POOL <- as.factor(shrimppool_all_bia$POOL)
glimpse(shrimppool_all_bia)
#let's make a wide dataset from shrimppool_all_bia that has year and
month in one column to use for later
W1shrimppool <- shrimppool_all_bia
W1shrimppool$Date <- paste(W1shrimppool$YEAR, W1shrimppool$Month, "1",
sep = "-")
W1shrimppool$Date <- ymd(W1shrimppool$Date)
wide_shrimppool <- W1shrimppool%>%
  group_by(Date, POOL, ATYACPUE, XIPHCPUE, MACCPUE)%>%
  select(-YEAR, -Month)
wide_shrimppool

#let's make a long dataset shrimppool_all_bia
#The species are also factors, let's make 2 columns:
```

```
#one for species as a factor varialbles
#one for the count values currently stored in the individual species
columns
T_shrimppool<- gather(shrimppool_all_bia,Species,Counts,-YEAR, -Month, -
POOL)
head(T_shrimppool)
tail(T_shrimppool)
T_shrimppool$Species <- as.factor(T_shrimppool$Species)
glimpse(T_shrimppool)
#Let's clean up the date. As no sample day was given, we will assume the
first of the month
#make a date out of the columns
T_shrimppool$Date <- paste(T_shrimppool$YEAR, T_shrimppool$Month, "1",
sep = "-")
glimpse(T_shrimppool)
head(T_shrimppool)
#format the date column
T_shrimppool$Date <- ymd(T_shrimppool$Date)
glimpse(T_shrimppool)
head(T_shrimppool)
#make another table that eliminates the YEAR and Month column
shrimppool_fin <- T_shrimppool %>%
  group_by(Date, POOL, Species)%>%
  select(-YEAR, -Month)
head(shrimppool_fin)
glimpse(shrimppool_fin)

####EDA-Visualizations to graphically understand data####
shrimppool_fin %>% arrange(shrimppool_fin$Date)
#scatterplot
ggplot(shrimppool_fin, aes(x = shrimppool_fin$Date, y =
shrimppool_fin$Counts, color=shrimppool_fin$Species))+
  geom_jitter(alpha = 0.6) +
  stat_smooth(method = "lm", se=FALSE, col = "red") +
  scale_y_continuous("Shrimp Species Found Per Trap") +
  scale_x_date("Year of Collection") +
  facet_grid(rows = vars(shrimppool_fin$POOL), cols =
vars(shrimppool_fin$Species)) +
  labs(title = "Shrimp Species Found by Date and Pool Location from
1988-2016", col = "Species")
#between the zero counts and low counts Maccpue sp. seems to show no
trends. let's plot on log scale and alone to see if more info is
revealed.
#log plot
ggplot(shrimppool_fin, aes(x = shrimppool_fin$Date, y =
shrimppool_fin$Counts, color=shrimppool_fin$Species))+
  geom_jitter(alpha = 0.6) +
  scale_y_log10("Shrimp Species Found Per Trap (log)") +
  scale_x_date("Year of Collection") +
  facet_grid(rows = vars(shrimppool_fin$POOL), cols =
vars(shrimppool_fin$Species)) +
```

```
  labs(title = "Shrimp Species Found by Date and Pool Location from
1988-2016", col = "Species")

#Maccpue sp. by date and location
maccpuectplot <- shrimppool_fin %>%
  filter(Species == "MACCPUE") %>%
  ggplot(aes(x = Date, y = Counts, color= POOL)) +
    geom_jitter(alpha = 0.6) +
    stat_smooth() +
    labs(x="Year of Collection", y="Maccpue sp. Found Per Trap", title =
"Maccpue sp. Found by Pool Location from 1988-2016", col = "Pool")
maccpuectplot

ggpairs(data = shrimppool_fin, columns = 1:4)
#The only evidence of a poor correlation is between the number of date
and the number of counts.
#Corr 0.135 particularly in the late 1990s and 2010s
#looks like the collection dates are inconsistent per pool and species.
ideally, this should be where i create a series of loops to compute the
average counts for each species biannually.
#i will come back to that; however, in the interest of time for this
assignment, i will leave the data as is.

#we want to see if there's correlation between species over time so
let's look at three more plots:
#calculate means of each species with all pools assummed equal
mean_Asp <- mean(wide_shrimppool$ATYACPUE)
mean_Asp #38.69164
mean_Xsp <- mean(wide_shrimppool$XIPHCPUE)
mean_Xsp #22.24613
mean_Msp <- mean(wide_shrimppool$MACCPUE)
mean_Msp #0.1786804
#calculate means of each species by pool
#Pool 0 by species
mean_bia0A <- shrimppool_all_bia %>%
  filter(POOL == '0') %>%
  summarize(mean0A = mean(shrimppool_all_bia$ATYACPUE), mean0X =
mean(shrimppool_all_bia$XIPHCPUE), mean0M =
mean(shrimppool_all_bia$MACCPUE))
#Pool 8 by species
mean_bia8A <- shrimppool_all_bia %>%
  filter(POOL == '8') %>%
  summarize(mean8A = mean(shrimppool_all_bia$ATYACPUE),mean8X =
mean(shrimppool_all_bia$XIPHCPUE), mean8M =
mean(shrimppool_all_bia$MACCPUE))
#Pool 15 by species
mean_bia15A <- shrimppool_all_bia %>%
  filter(POOL == '15') %>%
  summarize(totmean15A = mean(shrimppool_all_bia$ATYACPUE), mean15X =
mean(shrimppool_all_bia$XIPHCPUE), mean15M =
mean(shrimppool_all_bia$MACCPUE))
```

```
#use the wide dataset wide_shrimppool to compare ATYACPUE-XIPHCPUE,
ATYACPUE-MACCPUE, XIPHCPUE-MACCPUE across all pools
#as ATYACPUE had the largest average observations per pool, we will
assume it to be the independent variable (x) in all species comparisons.
#as MACCPUE had the smallest average observations per pool, we will
assume it to be the dependent variable (y) in all species comparisons.
#as the average observations per pool for XIPHCPUE were <ATYACPUE and
>MACCPUE, we will assume it to be the dependent variable (x) in
comparison to ATYACPUE,
#and the independent variable (x) in comparison to MACCPUE
#x=ATYACPUE-y=XIPHCPUE
ggplot(wide_shrimppool, aes(x=wide_shrimppool$ATYACPUE,
y=wide_shrimppool$XIPHCPUE, color=wide_shrimppool$POOL))+
  geom_point(alpha=0.6) +
  stat_smooth(method = "lm", se=FALSE, col = "red") +
  scale_y_continuous("Average Populations of Xiphocaris elongata") +
  scale_x_continuous("Average Populations of Atya lanipes") +
  facet_grid(wide_shrimppool$POOL) +
  labs(title = "Average Populations of Atya lanipes to Xiphocaris
elongata, 1988-2016", color="Pool")
#x=ATYACPUE-y=MACCPUE
ggplot(wide_shrimppool, aes(x=wide_shrimppool$ATYACPUE,
y=wide_shrimppool$MACCPUE, color=wide_shrimppool$POOL))+
  geom_point(alpha=0.6) +
  stat_smooth(method = "lm", se=FALSE, col = "red") +
  scale_y_continuous("Average Populations of Macrobrachium spp") +
  scale_x_continuous("Average Populations of Atya lanipes") +
  facet_grid(wide_shrimppool$POOL) +
  labs(title = "Average Populations of Atya lanipes to Macrobrachium
spp, 1988-2016", color="Pool")
#x=XIPHCPUE-y=MACCPUE
ggplot(wide_shrimppool, aes(x=wide_shrimppool$XIPHCPUE,
y=wide_shrimppool$MACCPUE, color=wide_shrimppool$POOL))+
  geom_point(alpha=0.6) +
  stat_smooth(method = "lm", se=FALSE, col = "red") +
  scale_y_continuous("Average Populations of Macrobrachium spp") +
  scale_x_continuous("Average Populations of Xiphocaris elongata") +
  facet_grid(wide_shrimppool$POOL) +
  labs(title = "Average Populations of Xiphocaris elongata to
Macrobrachium spp, 1988-2016", color="Pool")

head(wide_shrimppool)
#ggpairs(data = wide_shrimppool, columns = 1:5)
ggpairs(data = wide_shrimppool, columns = 2:5)

#standard deviation between species
by (wide_shrimppool$ATYACPUE, wide_shrimppool$XIPHCPUE,
wide_shrimppool$MACCPUE, sd)
#standard deviation between species and time
by (shrimppool_fin$Date, shrimppool_fin$Counts, sd)
```

```
####Linear Models####
#lm(y~x,data), where y is the dependent variable, x is the independent
variable
#create a unified wide dataframe with all original data, all explanatory
variables, and residuals
#use null model to find out how well our model performed

##ALL SPECIES TO TIME##
lm_spyr <- lm(Counts ~ Date, data=shrimppool_fin)
lm_spyr #intercept=9.117668; Date= 0.001019
summary (lm_spyr)
coef(lm_spyr)
#vector with all the fitted values (y'), which will tell us what the
model predicted
fitted_mx <- fitted.values(lm_spyr)
#residuals from fitted values, which tells the difference between the
actual, measured value and the predicted (fitted) values
res_spyr <- residuals(lm_spyr)
#Residuals:
#Min      1Q  Median      3Q     Max
#-26.555 -17.724  -5.415  10.595 131.431
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept) 9.117668   2.695591   3.382 0.000746 ***
#  Date        0.001019   0.000235   4.338 1.58e-05 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Residual standard error: 23.43 on 1021 degrees of freedom
#Multiple R-squared:  0.0181,   Adjusted R-squared:  0.01714
#F-statistic: 18.82 on 1 and 1021 DF,  p-value: 1.577e-05
#unified dataframe for all species to year
lm_spyr_un <- broom::augment(lm_spyr)
head(lm_spyr_un)
lm_spyr_un %>%
  arrange(desc(.resid))%>%
  head() %>%
  tail()
lm_spyr_un$.resid_abs<-abs(lm_spyr_un$.resid_abs)
lm_spyr_un %>%
  arrange(desc(.resid_abs)) %>%
  head()
#inspect outlier Date==1999-06-01
shrimppool_fin%>%
  filter(Date == '1999-06-01') #152 Atyacpue species in Pool 15. As the
other species were only counted
#at 13.5 and 0 in Pool 15 on that date, a question for further
examination would be what is causing Atya sp to proliferate.
#create null model
spyr_null <- lm(Counts ~1, data = shrimppool_fin)
```

```
spyr_null #intercept = 20.37
#verify null model
mean_spcount <- mean(shrimppool_fin$Counts)
mean_spcount #20.37215
ggplot(data = shrimppool_fin, aes(x=Date, y=Counts))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_spcount) +
  labs(title = "Species Count Null Model")
ggplot(data = shrimppool_fin, aes(x=Date, y=Counts))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+
  geom_hline(yintercept = mean_spcount) +
  labs(title = "Species Count Null Model")
#assess error = Multiple R-squared
summary(lm_spyr) #0.0181 which is not good as most points are outside
the the standard error so the model cannot accurately predict the amount
of species related to year


##XIPHCPUE-ATYACPUE##
lm_xa <- lm(XIPHCPUE ~ ATYACPUE, data=wide_shrimppool)
lm_xa #intercept= 13.268; ATYACPUE=0.232
summary (lm_xa)
coef(lm_xa)
#vector with all the fitted values (y'), which will tell us what the
model predicted
fitted_xa <- fitted.values(lm_xa)
#residuals from fitted values, which tells the difference between the
actual, measured value and the predicted (fitted) values
res_xa <- residuals(lm_xa)
#Residuals:
#  Min      1Q  Median      3Q     Max
#-34.921 -12.015  -3.749   8.552  85.810
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)  13.2685     1.7237   7.697 1.53e-13 ***
#  ATYACPUE      0.2320     0.0376   6.171 1.93e-09 ***
#  ---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Residual standard error: 17.08 on 339 degrees of freedom
#Multiple R-squared:  0.101,    Adjusted R-squared:  0.09835
#F-statistic: 38.09 on 1 and 339 DF,  p-value: 1.931e-09
#unified dataframe for all species to year
lm_xa_un <- broom::augment(lm_xa)
head(lm_xa_un)
lm_xa_un %>%
  arrange(desc(.resid))%>%
  head() %>%
  tail()
#inspect outlier XIPHCPUE==113 & ATYACPUE==60
wide_shrimppool %>%
```

```
  filter(XIPHCPUE==113 & ATYACPUE==60) #This outlier occured on 2015-11-
01. The  XIPHCPUE==113 > ATYACPUE==60
#Contrary to our calculated mean  XIPHCPUE==113 > ATYACPUE==60. What
caused this change in population dynamic?
#create null model
xa_null <- lm(XIPHCPUE~1, data = wide_shrimppool)
xa_null #intercept = 22.25
#verify null model
mean_Xsp
ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=XIPHCPUE))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_Xsp) +
  labs(title = "Xiphocaris elongate Species Count to Null Model")
ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=XIPHCPUE))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+
  geom_hline(yintercept = mean_Xsp) +
  labs(title = "Xiphocaris elongate Species Count to Null Model")
#assess error = Multiple R-squared
summary(lm_xa) #0.0181 which is not good as most points are outside the
the standard error so the model cannot accurately predict the amount of
species related to year


##MACCPUE-ATYACPUE##
lm_ma <- lm(MACCPUE~ATYACPUE, data=wide_shrimppool)
lm_ma #intercept=0.207005614; ATYACPUE=-0.000732077
summary (lm_ma)
coef(lm_ma)
#vector with all the fitted values (y'), which will tell us what the
model predicted
fitted_ma <- fitted.values(lm_ma)
#residuals from fitted values, which tells the difference between the
actual, measured value and the predicted (fitted) values
res_ma <- residuals(lm_ma)
#Residuals:
#Min      1Q  Median      3Q     Max
#-0.2070 -0.1831 -0.1382  0.1128  1.5113
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)  0.2070056  0.0291442   7.103 7.23e-12 ***
#  ATYACPUE    -0.0007321  0.0006357  -1.152     0.25
#---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Residual standard error: 0.2887 on 339 degrees of freedom
#Multiple R-squared:  0.003897, Adjusted R-squared:  0.0009587
#F-statistic: 1.326 on 1 and 339 DF,  p-value: 0.2503
lm_ma_un <- broom::augment(lm_ma)
head(lm_ma_un)
lm_ma_un %>%
  arrange(desc(.resid))%>%
```

```
  tail()
#inspect outlier MACCPUE==1.7 & ATYACPUE==25
wide_shrimppool %>%
  filter(MACCPUE==1.7 & ATYACPUE==25) #This outlier occured on 1996-06-
01.What caused this small count in Atyacpue in Pool 8?
#inspect outlier MACCPUE==0 & ATYACPUE==6
wide_shrimppool %>%
  filter(MACCPUE==0 & ATYACPUE==6) #This outlier occured on 1989-12-01.
What caused this small count for all species in Pool 8?
#create null model
ma_null <- lm(MACCPUE~1, data = wide_shrimppool)
ma_null #intercept = 0.1787
#verify null model
mean_Msp
ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=MACCPUE))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium spp Species Count to Null Model")
ggplot(data = wide_shrimppool, aes(x=ATYACPUE, y=MACCPUE))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium Species Count to Null Model")
#assess error = Multiple R-squared
summary(lm_ma) #0.029  the species are not correlated


##MACCPUE-XIPHCPUE##
lm_mx <- lm(MACCPUE~XIPHCPUE, data=wide_shrimppool)
lm_mx #intercept=0.2141961; XIPHCPUE=-0.0015965
summary (lm_mx)
coef(lm_mx)
#vector with all the fitted values (y'), which will tell us what the
model predicted
fitted_mx <- fitted.values(lm_mx)
#residuals from fitted values, which tells the difference between the
actual, measured value and the predicted (fitted) values
res_mx <- residuals(lm_mx)
#Residuals:
#  Min      1Q  Median      3Q     Max
#-0.2142 -0.1914 -0.1299  0.1090  1.5013
#Coefficients:
#  Estimate Std. Error t value Pr(>|t|)
#(Intercept)  0.2141961  0.0248170   8.631  2.4e-16 ***
#  XIPHCPUE    -0.0015965  0.0008681  -1.839   0.0668 .
#---
#  Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
#Residual standard error: 0.2878 on 339 degrees of freedom
#Multiple R-squared:  0.009879, Adjusted R-squared:  0.006958
#F-statistic: 3.382 on 1 and 339 DF,  p-value: 0.06677
lm_mx_un <- broom::augment(lm_mx)
```

```
head(lm_mx_un)
lm_mx_un %>%
  arrange(desc(.resid))%>%
  head() %>%
  tail()
#inspect outlier MACCPUE==0.31 & Xiphcpue==8.25
wide_shrimppool %>%
  filter(MACCPUE==0.31 & XIPHCPUE==8.25) #This outlier occured on 1988-
01-01 in Pool0.What caused the counts of A&X to be nearly equal?
#inspect outlier MACCPUE==1.7 & Xiphcpue==9.7
wide_shrimppool %>%
  filter(MACCPUE==1.7 & XIPHCPUE==9.7) #This outlier occured in Pool 8
on  1996-06-01.
#create null model
mx_null <- lm(MACCPUE~1, data = wide_shrimppool)
mx_null #intercept = 0.1787
#verify null model
mean_Msp
ggplot(data = wide_shrimppool, aes(x=XIPHCPUE, y=MACCPUE))+
  geom_point(alpha=0.6)+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium spp Species Count to Null Model")
ggplot(data = wide_shrimppool, aes(x=XIPHCPUE, y=MACCPUE))+
  geom_point(alpha=0.6)+
  stat_smooth(method = "lm")+
  geom_hline(yintercept = mean_Msp) +
  labs(title = "Maccrobrachium Species Count to Null Model")
#assess error = Multiple R-squared
summary(lm_mx) #0.009879 which is a horizontal line indicating the
species are not correlated
```