R Homework

Heva, Prita, Supria

April 6, 2018

Data Visualization in R : Long-term monitoring of a rodent community

Our homework is analyzing database of a long term monitoring for rodent community in Chihuahuan Desert ecosystem near Portal, Arizona, from 1977 to 2000. At this site, 24 experimental plots were established in 1977 and divided among controls and experimental manipulations. The long-term data for the rodent community at the Portal Project has been used to address a variety of questions including:

1. Monitoring the population-level dynamics of desert rodents & competitive interactions among rodent species.
2. Responses of rodents to climatic variability.
3. The long-term stability and dynamics of a desert rodent community.

This is our R-homework documentation. The process of data visualisation using R can be divided into four steps:

- Load the library
- Read the data file
- Clean the data
- Analyze the data
    1. Time series data of sex and number of sample per plot type
    2. Correlation between length of hindfoot and weight of animal
    3. Changes of weight over the year based on each plot type
    4. Changes of length of hindfoot over the year based on each plot type
    5. The relationship between hindfoot_length and weight on each plot type
    6. Correlation between hindfoot length and genus

Load the library

```
library(tidyverse)
library(lubridate)
library(gridExtra)
library(ggplot2)
library(dbplyr)
library(ggpubr)
library(kableExtra)
```

Read the data file

Our team decide to read the combined.csv file because it has the most comprehensive, consise, and compact data.

```
surveys_combined <- read.csv("data/combined.csv")
```

Below is the information about the data structure:

```
#find the unique value of observed columns
csex <- toString(levels(unique(surveys_combined$sex)))
cplot_type <- toString(levels(unique(surveys_combined$plot_type)))
cplot_id <- toString(sort(unique(surveys_combined$plot_id), decreasing = FALSE))
cgenus <- toString(levels(unique(surveys_combined$genus)))
cspecies <- toString(levels(unique(surveys_combined$species)))
cspecies_id <- toString(levels(unique(surveys_combined$species_id)))
#create the table of unique value
tbl_str <- data.frame( "Sex" = c(csex), "Plot Type" = c(cplot_type),
                        "Plot ID" = c(cplot_id), "Genus" = c(cgenus),
                        "Species" = c(cspecies), "Species ID" = c(cspecies_id))
kable(tbl_str, "html") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

Sex	Plot.Type	Plot.ID	Genus	Species	Species.ID
, F, M	Control, Long-term Krat Exclosure, Rodent Exclosure, Short-term Krat Exclosure, Spectab exclosure	1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	Ammodramus, Ammospermophilus, Amphispiza, Baiomys, Calamospiza, Callipepla, Campylorhynchus, Chaetodipus, Cnemidophorus, Crotalus, Dipodomys, Lizard, Neotoma, Onychomys, Perognathus, Peromyscus, Pipilo, Pooecetes, Reithrodontomys, Rodent, Sceloporus, Sigmodon, Sparrow, Spermophilus, Sylvilagus, Zonotrichia	albigula, audubonii, baileyi, bilineata, brunneicapillus, chlorurus, clarki, eremicus, flavus, fulvescens, fulviventer, fuscus, gramineus, harrisi, hispidus, intermedius, leucogaster, leucophrys, leucopus, maniculatus, megalotis, melanocorys, merriami, montanus, ochrognathus, ordii, penicillatus, savannarum, scutalatus, sp., spectabilis, spilosoma, squamata, taylori, tereticaudus, tigris, torridus, undulatus, uniparens, viridis	AB, AH, AS, BA, CB, CM, CQ, CS, CT, CU, CV, DM, DO, DS, DX, NL, OL, OT, OX, PB, PC, PE, PF, PG, PH, PI, PL, PM, PP, PU, PX, RF, RM, RO, RX, SA, SC, SF, SH, SO, SS, ST, SU, UL, UP, UR, US, ZL

Clean the data

Our team read the raw data and transform it into consistent data that can be analyzed. It is aimed at improving the content of statistical statements based on the data as well as their reliability. This proces is *data cleaning*. In this homework, and we ignore the missing data *(”, NULL, is.Na)*.

```
surveys_combined_clear<- surveys_combined %>%
  filter(!is.na(sex), sex != "", !is.na(hindfoot_length),hindfoot_length != "",
         !is.na(weight), weight != "")
```

The result of data cleaning :

```
#create the table of nrow value before and after cleaning the data
tbl_nrow <- data.frame( "Before" = prettyNum(nrow(surveys_combined), big.mark = ","),
                        "After" = prettyNum(nrow(surveys_combined_clear), big.mark = ","))
kable(tbl_nrow, "html") %>%
  kable_styling(
    bootstrap_options = c("striped", "hover", "condensed", "responsive"), full_width = F) %>%
    add_header_above(c("Data Cleaning " = 2))
```

Data Cleaning	
Before	After
34,786	30,676

For the simple distribution tables below the 1st and 3rd Qu. refer to the first and third quartiles, indicating that 25% of the observations have values of that variable which are less than or greater than (respectively) the value listed.

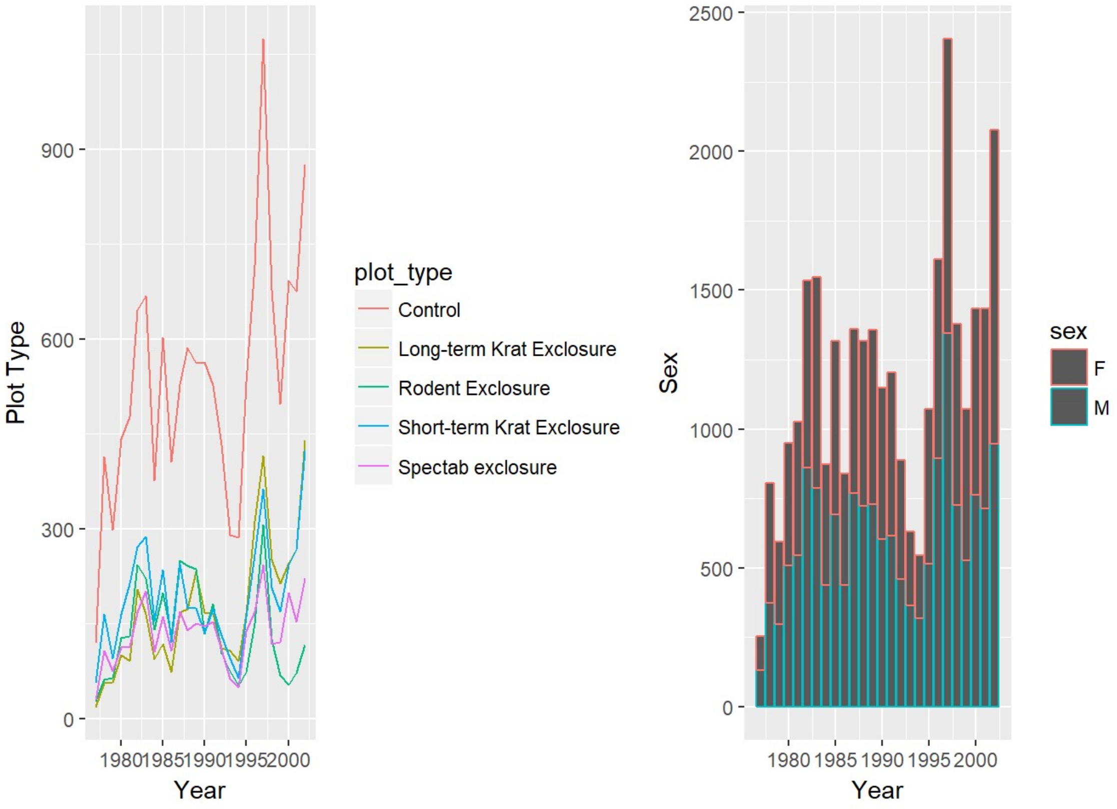
```
#create statatistical summarize
summary(surveys_combined_clear)
```

```
##    record_id         month             day             year     
##  Min.   :   63   Min.   : 1.000   Min.   : 1.00   Min.   :1977  
##  1st Qu.: 9882   1st Qu.: 4.000   1st Qu.: 9.00   1st Qu.:1985  
##  Median :18659   Median : 7.000   Median :16.00   Median :1991  
##  Mean   :18475   Mean   : 6.552   Mean   :16.05   Mean   :1991  
##  3rd Qu.:27132   3rd Qu.:10.000   3rd Qu.:23.00   3rd Qu.:1997  
##  Max.   :35548   Max.   :12.000   Max.   :31.00   Max.   :2002  
##                                                                 
##     plot_id        species_id   sex       hindfoot_length     weight      
##  Min.   : 1.00   DM     :9727    :    0   Min.   : 2.00   Min.   :  4.00  
##  1st Qu.: 5.00   PP     :2969   F:14584   1st Qu.:21.00   1st Qu.: 20.00  
##  Median :11.00   PB     :2803   M:16092   Median :31.00   Median : 36.00  
##  Mean   :11.22   DO     :2790             Mean   :29.21   Mean   : 41.79  
##  3rd Qu.:17.00   RM     :2417             3rd Qu.:36.00   3rd Qu.: 47.00  
##  Max.   :24.00   OT     :2081             Max.   :64.00   Max.   :280.00  
##                  (Other):7889                                             
##              genus               species          taxa      
##  Dipodomys      :14540   merriami    :9727   Bird   :    0  
##  Chaetodipus    : 5781   penicillatus:2969   Rabbit :    0  
##  Onychomys      : 2991   baileyi     :2803   Reptile:    0  
##  Reithrodontomys: 2500   ordii       :2790   Rodent :30676  
##  Peromyscus     : 2068   megalotis   :2417                  
##  Perognathus    : 1500   torridus    :2081                  
##  (Other)        : 1296   (Other)     :7889                  
##                      plot_type    
##  Control                  :13972  
##  Long-term Krat Exclosure : 4517  
##  Rodent Exclosure         : 3595  
##  Short-term Krat Exclosure: 5062  
##  Spectab exclosure        : 3530  
##                                   
## 
```

Analyze the data

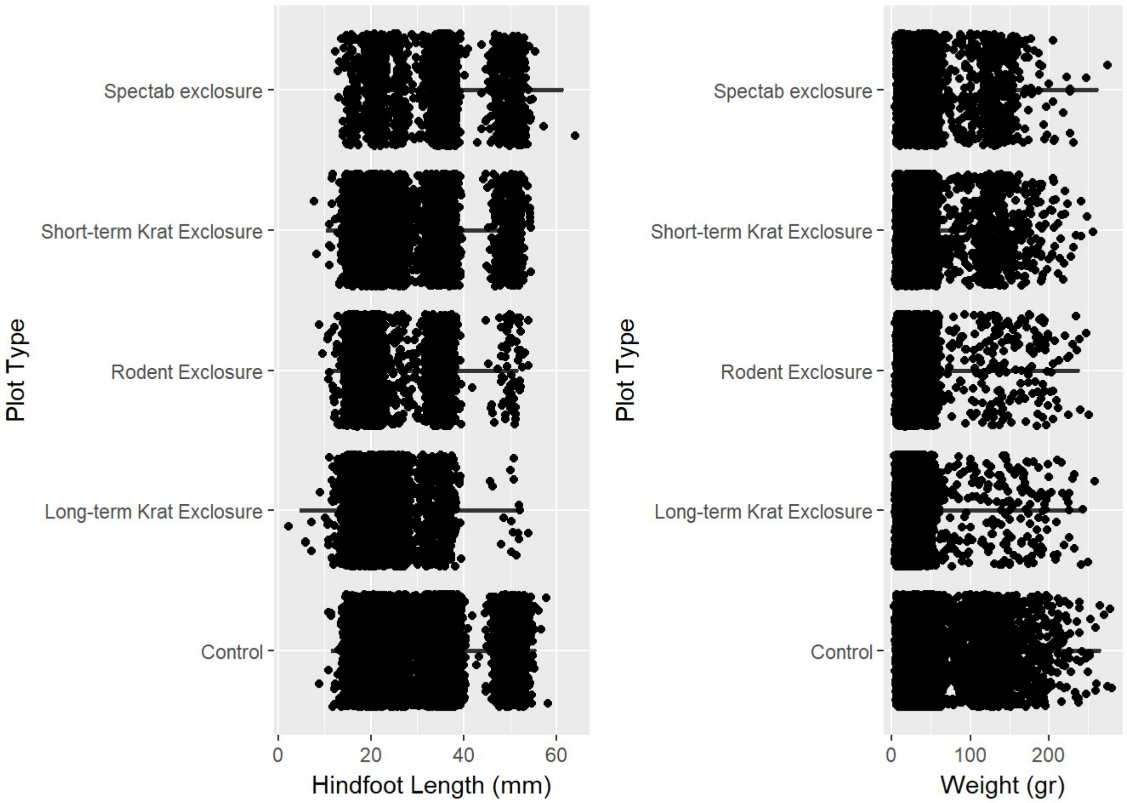
1. Time series data of sex and number of sample per plot type

```
#create line chart plot type per year
year_plot_type <- surveys_combined_clear %>% group_by(year, plot_type) %>% tally()
line_chart <- ggplot(year_plot_type, aes(x=year, y=n, color=plot_type)) +
  geom_line() + xlab("Year") + ylab("Plot Type")
#create bar chart sex per year
year_sex <- surveys_combined_clear %>% group_by(year, sex) %>% tally()
bar_chart <- ggplot(year_sex, aes(x=year, y=n, color=sex)) +
  geom_bar(stat="identity") + xlab("Year") + ylab("Sex")
#put chart to grid
timeseries_plot <- grid.arrange(line_chart, bar_chart, ncol=2, widths=c(8,6))
```



The line graph illustrates the number of rodent sample among controls and other experimental manipulations. Each sample indicates data for every desert rodent caught on the 20 ha. And the bar graph shows the number of rodent differentiate by sex. Overall, there is a trend of decreasing number of sample but it increase in 2000.

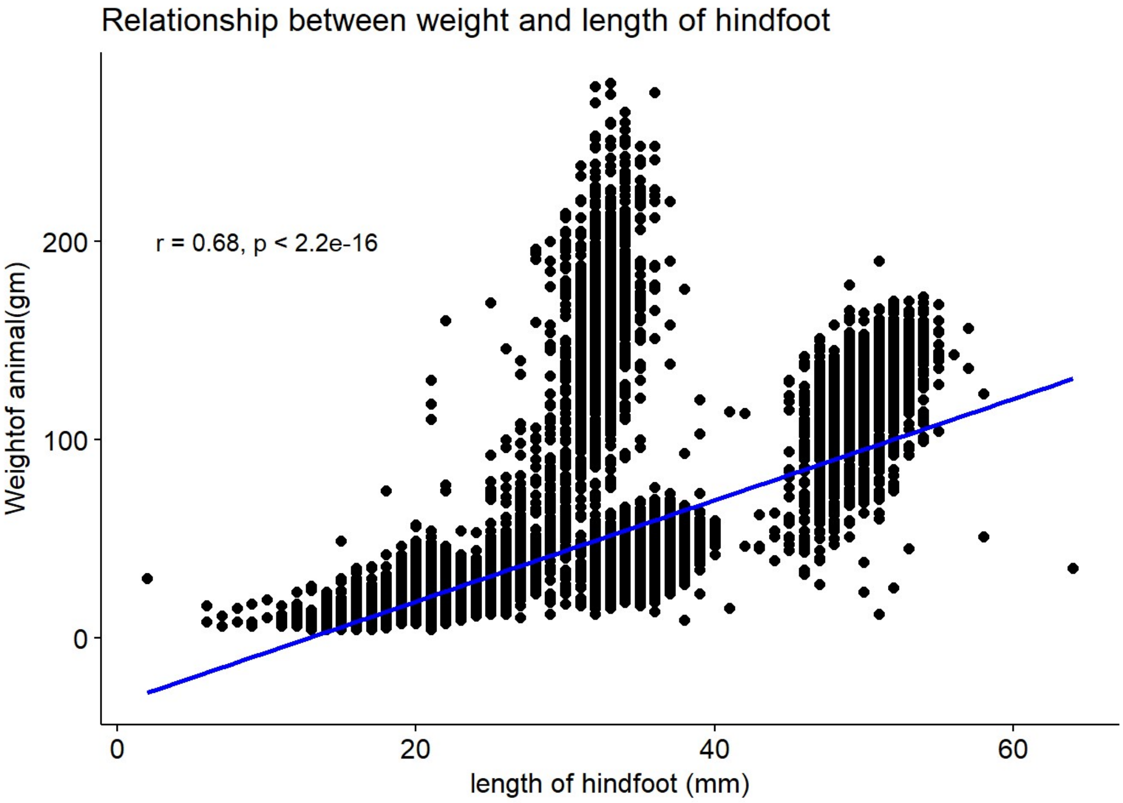
```
#create boxplot chart weight per plot_type
boxplot_chart_weight <- ggplot(surveys_combined_clear, aes(x=weight, y=plot_type))+
  geom_boxplot()+xlab("Weight (gr)")+ylab("Plot Type") +  geom_jitter()
#create boxplot chart hindfoot length per plot_type
boxplot_chart_length <- ggplot(surveys_combined_clear, aes(x=hindfoot_length, y=plot_type))+
  geom_boxplot()+xlab("Hindfoot Length (mm)")+ylab("Plot Type") + geom_jitter()
#put chart to grid
frequency_plot <- grid.arrange(boxplot_chart_length, boxplot_chart_weight, ncol=2, widths=c(8,7))
```



This boxplot graph explain the distribution of hindfoot length and weight per plot type. Detail correlation will be explain below.

2. Correlation between length of hindfoot and weight of animal

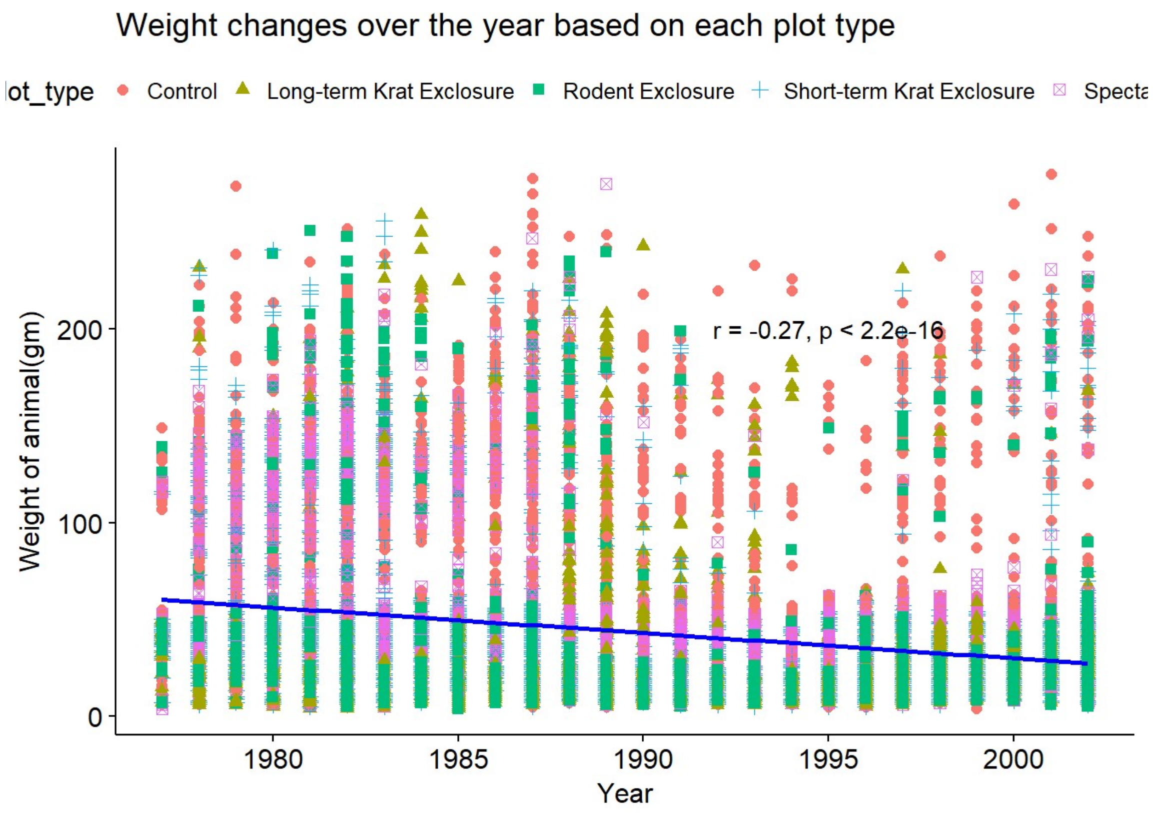
```
# Created by supria
# Scatter plot with correlation coefficient
#:::::::::::::::::::::::::::::::::::::::::::::::::
sp <- ggscatter(surveys_combined_clear, x = "hindfoot_length", y = "weight",
                title = "Relationship between weight and length of hindfoot",
                xlab = "length of hindfoot (mm)", ylab = "Weightof animal(gm)",
                add = "reg.line",  # Add regressin line
                add.params = list(color = "blue", fill = "lightgray"), # Customize reg. line
                conf.int = TRUE # Add confidence interval
                )
# Add correlation coefficient
final.plot <- sp + stat_cor(method = "pearson", label.x = 10, label.y = 200)
final.plot
```



This plot means that what is the relationship between weight of animal and length of hindfoot length. Using statistic analysis we found that there is a linear correlation. And the value of R-squared is greater than .5. It has shown there is 68% linearly correlated.

3. Changes of weight over the year based on each plot type

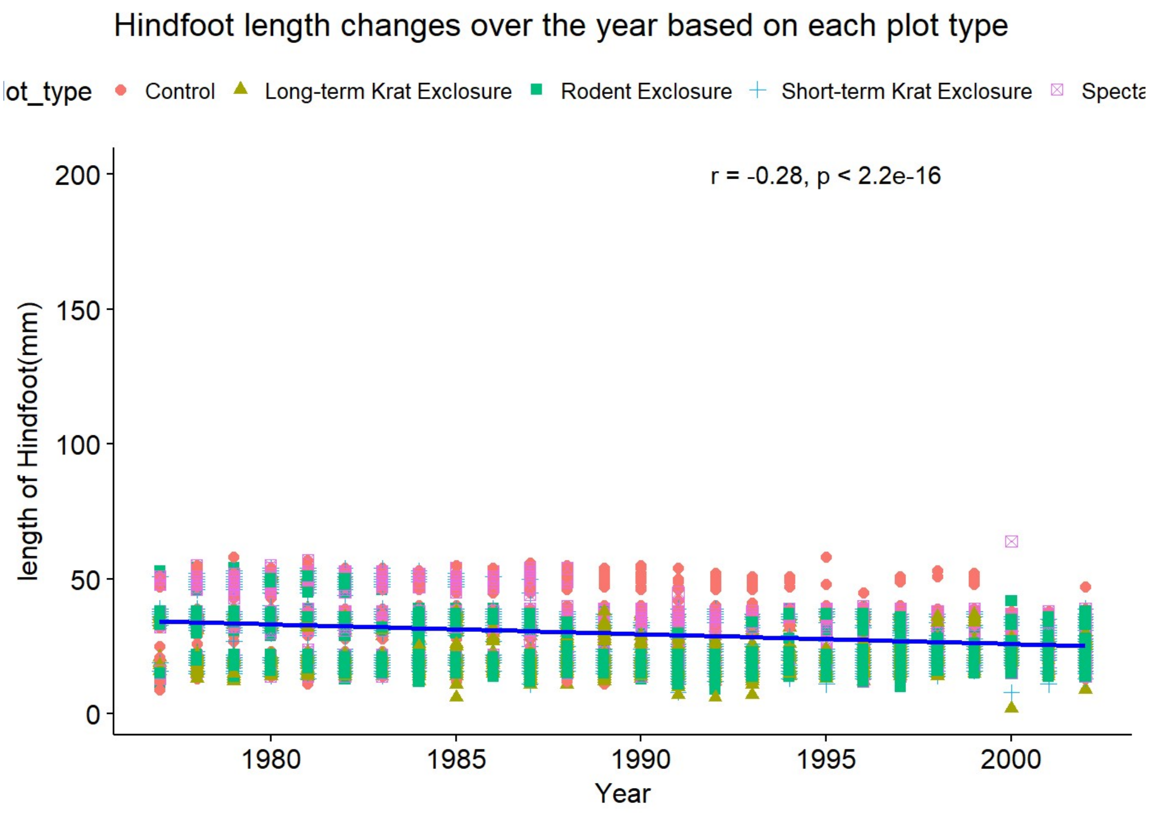
```
# created by supria
wg <- ggscatter(data=surveys_combined_clear, x='year',y='weight',
                color ="plot_type", shape = "plot_type",
                title = "Weight changes over the year based on each plot type",
                xlab = "Year", ylab = "Weight of animal(gm)",
                add = "reg.line",  # Add regressin line
                add.params = list(color = "blue","red","green","yellow","pink",
                                  fill = "lightgray"), # Customize reg. line
                conf.int = TRUE # Add confidence interval
                )
# Add correlation coefficient
final.wg <- wg + stat_cor(method = "pearson", label.x = 1995, label.y = 200)
final.wg
```



This plot explains about what is the change of weight over the year for each plot type It has shown that there is no relationship of weight over the year. R-squared value explains that -.27 which is very low.

4. Changes of length of hindfoot over the year based on each plot type

```
# Created by supria
# Extend the regression lines beyond the domain of the data
##hindfoot changes over the year
hd <- ggscatter(data=surveys_combined_clear, x='year',y='hindfoot_length',
                color ="plot_type",shape = "plot_type",
                title = "Hindfoot length changes over the year based on each plot type",
                xlab = "Year", ylab = "length of Hindfoot(mm)",
                add = "reg.line",  # Add regressin line
                add.params = list(color = "blue","red","green","yellow","pink",
                                  fill = "lightgray"), # Customize reg. line
                conf.int = TRUE # Add confidence interval
                )
# Add correlation coefficient
final.hd<- hd + stat_cor(method = "pearson", label.x = 1995, label.y = 200)
final.hd
```



This plot explains about what is the change of length of hindfoot over the year for each plot type. It has shown that there is no relationship of weight over the year. R-squared value explains that -.28 which is very low.

5. The relationship between hindfoot_length and weight on each plot type

We examined the relationship of the weight and hindfoot_length each plot type; control, long-term, rodent, short-term, and spectab exclosure. We tried to find the evident whether the increase of hindfoot_length every mm will contribute the increase or the decrease of the weight (gram) in every plot_type, and whether there is any significant difference or not among plot type. However, we compared it in genus level, not in species_id level, because we wanted to know whether there is any correlation of the chosen plot_type with the size of the weight and hindfoot_length. We also checked the r square and p value to understand about the variances and correlation.

```
#get hindfoot_length, weight, genus, plot type per each plot type
lw_control <- surveys_combined_clear %>% select(hindfoot_length, weight , genus, plot_type) %>%
  filter(plot_type == "Control")
lw_longterm <- surveys_combined_clear %>% select(hindfoot_length, weight , genus, plot_type) %>%
  filter(plot_type == "Long-term Krat Exclosure")
lw_rodent <- surveys_combined_clear %>% select(hindfoot_length, weight , genus, plot_type) %>%
  filter(plot_type == "Rodent Exclosure")
lw_shortterm <- surveys_combined_clear %>% select(hindfoot_length, weight , genus, plot_type) %>%
  filter(plot_type == "Short-term Krat Exclosure")
lw_spectab <- surveys_combined_clear %>% select(hindfoot_length, weight , genus, plot_type) %>%
  filter(plot_type == "Spectab exclosure")
#create plot & examine the correlation each plot type and hindfootlength
par(mfrow=c(3,2))
lw_control_plot <- plot(lw_control$weight, lw_control$hindfoot_length, xlab = "Weight (gr)",
                        ylab = "Plot Type (mm)", main = "Control")
lw_longterm_plot <- plot(lw_longterm$weight, lw_longterm$hindfoot_length, xlab = "Weight (gr)",
                        ylab = "Plot Type (mm)", main = "Long-term Krat Exclosure")
lw_rodent_plot <- plot(lw_rodent$weight, lw_rodent$hindfoot_length, xlab = "Weight (gr)",
                        ylab = "Plot Type (mm)", main = "Rodent Exclosure")
lw_shortterm_plot <- plot(lw_shortterm$weight, lw_shortterm$hindfoot_length,
                          xlab = "Weight (gr)", ylab = "Plot Type (mm)",
                          main = "Short-term Krat Exclosure")
lw_spectab_plot <- plot(lw_spectab$weight, lw_spectab$hindfoot_length, xlab = "Weight (gr)",
                        ylab = "Plot Type (mm)", main = "Spectab exclosure")
```

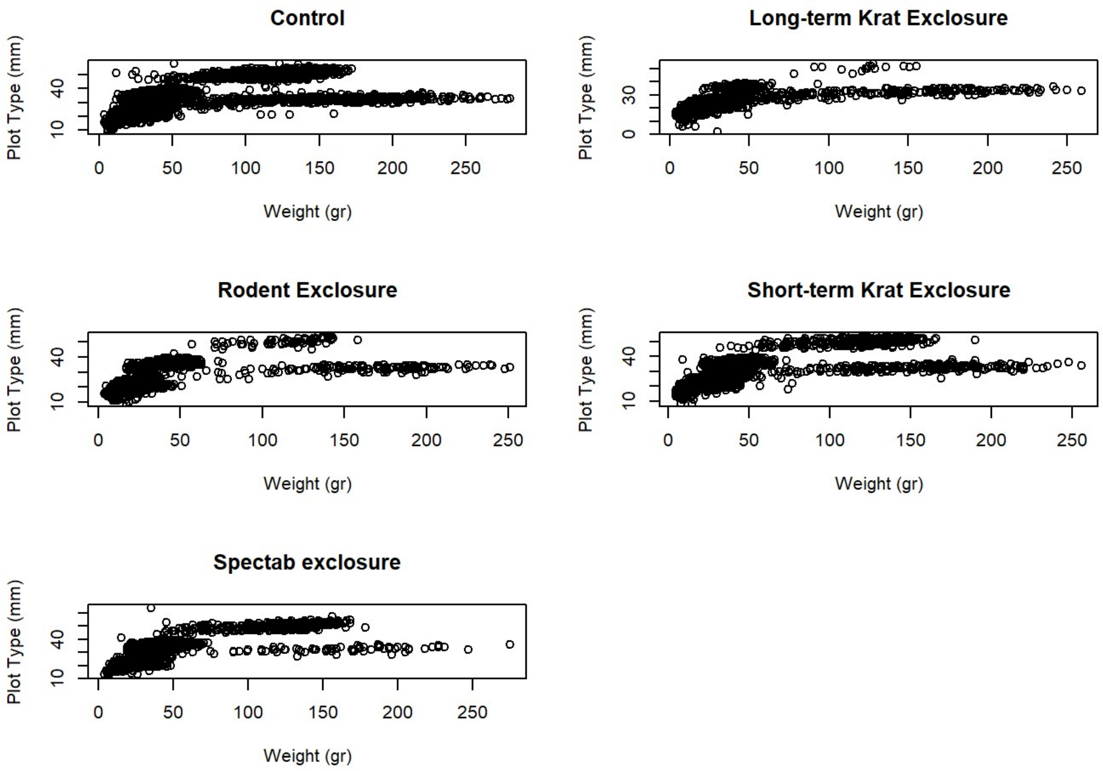


Table below show the relationship between hindfoot_length and weight on each plot type

```
#get a correlation value per plot type
lw_control_cor <- cor(lw_control$weight, lw_control$hindfoot_length)
lw_longterm_cor <- cor(lw_longterm$weight, lw_longterm$hindfoot_length)
lw_rodent_cor <- cor(lw_rodent$weight, lw_rodent$hindfoot_length)
lw_shortterm_cor <- cor(lw_shortterm$weight, lw_shortterm$hindfoot_length)
lw_spectab_cor <- cor(lw_spectab$weight, lw_spectab$hindfoot_length)
#create data frame for correlation
tbl_cor <- data.frame("Control" = c(lw_control_cor), "Long Term" = c(lw_longterm_cor),
                      "Rodent" = c(lw_rodent_cor), "Short Term" = c(lw_shortterm_cor),
                      "Spectab" = c(lw_spectab_cor))
#get fit linear model (lm)
lw_control_lm <- lm(lw_control$hindfoot_length ~ lw_control$weight)
lw_longterm_lm <- lm(lw_longterm$hindfoot_length ~ lw_longterm$weight)
lw_rodent_lm <- lm(lw_rodent$hindfoot_length ~ lw_rodent$weight)
lw_shortterm_lm <- lm(lw_shortterm$hindfoot_length ~ lw_shortterm$weight)
lw_spectab_lm <- lm(lw_spectab$hindfoot_length ~ lw_spectab$weight)
#get summary of fit linear model
lw_control_stat <- summary(lw_control_lm)
lw_longterm_stat <- summary(lw_longterm_lm)
lw_rodent_stat <- summary(lw_rodent_lm)
lw_shortterm_stat <- summary(lw_shortterm_lm)
lw_spectab_stat <- summary(lw_spectab_lm)
#get r square
lw_control_r <- lw_control_stat$r.squared
lw_longterm_r <- lw_longterm_stat$r.squared
lw_rodent_r <- lw_rodent_stat$r.squared
lw_shortterm_r <- lw_shortterm_stat$r.squared
lw_spectab_r <- lw_spectab_stat$r.squared
#create data frame for r square
tbl_r <- data.frame("Control" = c(lw_control_r), "Long Term" = c(lw_longterm_r),
                    "Rodent" = c(lw_rodent_r), "Short Term" = c(lw_shortterm_r),
                    "Spectab" = c(lw_spectab_r))
#get estimate weight
lw_control_ew <- lw_control_stat$coefficients["lw_control$weight","Estimate"]
lw_longterm_ew <- lw_longterm_stat$coefficients["lw_longterm$weight","Estimate"]
lw_rodent_ew <- lw_rodent_stat$coefficients["lw_rodent$weight","Estimate"]
lw_shortterm_ew <- lw_shortterm_stat$coefficients["lw_shortterm$weight","Estimate"]
lw_spectab_ew <- lw_spectab_stat$coefficients["lw_spectab$weight","Estimate"]
#create data frame for estimate weight
tbl_ew <- data.frame("Control" = c(lw_control_ew), "Long Term" = c(lw_longterm_ew),
                    "Rodent" = c(lw_rodent_ew), "Short Term" = c(lw_shortterm_ew),
                    "Spectab" = c(lw_spectab_ew))
#get standard error
#std.error <- modelCoeffs["speed", "Std. Error"]  # get std.error for speed
lw_control_se <- lw_control_stat$coefficients["lw_control$weight","Std. Error"]
lw_longterm_se <- lw_longterm_stat$coefficients["lw_longterm$weight","Std. Error"]
lw_rodent_se <- lw_rodent_stat$coefficients["lw_rodent$weight","Std. Error"]
lw_shortterm_se <- lw_shortterm_stat$coefficients["lw_shortterm$weight","Std. Error"]
lw_spectab_se <- lw_spectab_stat$coefficients["lw_spectab$weight","Std. Error"]
#get t-value
lw_control_t <- lw_control_ew/lw_control_se
lw_longterm_t <- lw_longterm_ew/lw_longterm_se
lw_rodent_t <- lw_rodent_ew/lw_rodent_se
lw_shortterm_t <- lw_shortterm_ew/lw_shortterm_se
lw_spectab_t <- lw_spectab_ew/lw_spectab_se
tbl_t <- data.frame("Control" = c(lw_control_t), "Long Term" = c(lw_longterm_t),
                    "Rodent" = c(lw_rodent_t), "Short Term" = c(lw_shortterm_t),
                    "Spectab" = c(lw_spectab_t))
#get p-value
lw_control_p <- formatC(2*pt(-abs(lw_control_t), length(lw_control)-1), format="e", digits=2)
lw_longterm_p <- formatC(2*pt(-abs(lw_longterm_t), length(lw_longterm)-1),format="e", digits=2)
lw_rodent_p <- formatC(2*pt(-abs(lw_rodent_t), length(lw_rodent)-1), format="e", digits=2)
lw_shortterm_p <-formatC(2*pt(-abs(lw_shortterm_t),length(lw_shortterm)-1),format="e",digits=2)
lw_spectab_p <- formatC(2*pt(-abs(lw_spectab_t), length(lw_spectab)-1), format="e", digits=2)
tbl_p <- data.frame("Control" = c(lw_control_p), "Long Term" = c(lw_longterm_p),
                    "Rodent" = c(lw_rodent_p), "Short Term" = c(lw_shortterm_p),
                    "Spectab" = c(lw_spectab_p))

#bind data
tbl_rbind <- rbind(tbl_cor, tbl_ew, tbl_r, tbl_p, tbl_t)
tbl_item <- data.frame(Values = c("Correlation", "Est. Weight", "R Square",
                                  "p Value", "t Value"))
tbl_cbind <- cbind(tbl_item, tbl_rbind)
#create table
kable(tbl_cbind, "html") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"))
```

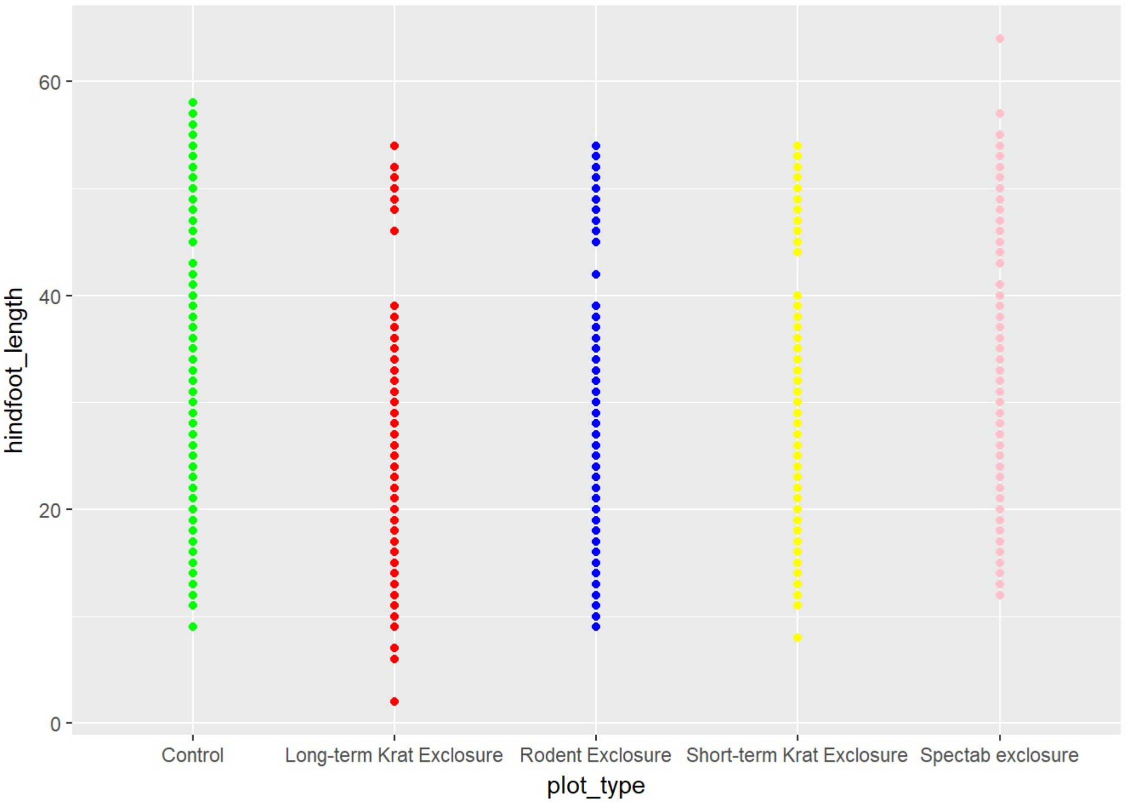
Values	Control	Long.Term	Rodent	Short.Term	Spectab
Correlation	0.639493347253542	0.651944411870459	0.637831402493435	0.713539815548686	0.732960644976775
Est. Weight	0.159900242351739	0.134576258646104	0.169057938435274	0.18186439354843	0.184122074435083
R Square	0.408951741181543	0.425031516169119	0.406828898006738	0.509139068373253	0.537231307084772
p Value	2.32e-06	1.14e-05	1.80e-05	5.80e-06	8.41e-06
t Value	98.3156744239618	57.7720029080713	49.6414562566589	72.4459708085272	63.997492711782

The table summaries the results above. The hindfoot_length is dependent variable and the weight is independent variable. For instance, it can be seen that the Spectab plot type has the greatest correlation of others, and the spectab has the higher correlation than control plot type (0.7327848 compare to 0.6397566). Like others, the Spectab also has positive and biggest weight (0.184094). It means that under the Spectab plot, every 1 mm increases of hindfoot_length will “increase” 0.184094 gram of the weight. The Spectab has the highest R Square number than others (0.537) which means that the Spectab explains 53,70% variances in the model. It has bigger variances which means that the values vary and disperse. The values do not congregate close to fitted line (mean) because the values vary (not relatively same) from mean value. p value is under 0.05 and very low which means Reject NULL hypothesis. Therefore, it has strong evidence of hindfoot_length and weight relationship as aforementioned.

6. Comparation between hindfoot length under each plot type

Plot 7 is supposed to compare among datasets. It is continuation from plot 6. After knowing the correlation, variance, and evidence levels, we want to compare 5 plot types to determine the best fit model.

```r
#Check the relationship and dispersal each plot
        length_control <- surveys_combined_clear%>%
          select(hindfoot_length, plot_type , genus) %>%
          filter(plot_type == "Control")

        length_Longterm <- surveys_combined_clear %>%
          select(hindfoot_length, plot_type , genus) %>%
          filter(plot_type == "Long-term Krat Exclosure")

        length_rodent <- surveys_combined_clear %>%
          select(hindfoot_length, plot_type , genus) %>%
          filter(plot_type == "Rodent Exclosure")

        length_shortterm <- surveys_combined_clear %>%
          select(hindfoot_length, plot_type , genus) %>%
          filter(plot_type == "Short-term Krat Exclosure")

        length_spectab <- surveys_combined_clear %>%
          select(hindfoot_length, plot_type , genus) %>%
          filter(plot_type == "Spectab exclosure")

        #Ggplot hindfoot_length and plot_type
          ggplot() +
          geom_point(data = length_control, aes(x=plot_type, y=hindfoot_length),
                     color = 'green') +
          geom_point(data = length_Longterm, aes(x=plot_type, y=hindfoot_length),
                     color = 'red') +
          geom_point(data = length_rodent, aes(x=plot_type, y=hindfoot_length),
                     color = 'blue') +
          geom_point(data = length_shortterm, aes(x=plot_type, y=hindfoot_length),
                     color = 'yellow') +
          geom_point(data = length_spectab, aes(x=plot_type, y=hindfoot_length),
                     color = 'pink')
```



To compare 5 variables should be done with Anova. However, it is challenging to combine 5 datasets and arrange them in associated columns. Nonetheless, from given graph, it looks like control and spectab resulting higher hindfoot_length and the resf of them; Long-term, short-term and rodent exclosure are relatively same. Hence, depending on purpose, we could select which methods we would use.