Visualizacion de datos en Lizard

Karen Huentenao Silva

16 de septiembre de 2018

Introduccion

Los libros de bioestadistica nos muestran como podemos sacar el mejor provecho de nuestros datos y dependiendo de su composicion que tipo de analisis o graficos son los que corresponden aplicar (Berke 2006)(Zar and others 1999). El trabajo en base de datos va a depender del tipo de datos que tengamos como de la pregunta a evaluar. Sin embargo en paper publicados en diferentes revistas es comun ver graficos que no son del todo adecuados para mostrar los resultados. Si bien el mensaje que se quiere mostrar queda claro al lector, siguen existiendo mejores formas de mostrar y presentar los resultados. Entonces en que basarnos para elegir el grafico.

Uno de los principales objetivos de este reporte mas que la parte biologica de los datos con los que se esta tabajando es realizar una comparacion en formas de presentar datos y realizar una exploracion mediante modelos aplicados.

Metodos

Este trabajo fue realizado basado en una base de datos de alometria en lagartos (Meiri 2010) con 915 observaciones. Se utilizo el programa R. Para poder trabajar en R y realizar una exploracion de datos se utilizaron diferentes de paquetes como el tidyverse (Wickham 2017), dismo, tidyr (**???**) dplyr(Wickham and others 2014), rmarkdown (Allaire et al., n.d.), knirt(Xie 2015), stargazer(Hlavac 2018). En particular para los graficos se utilizo el paquete ggplot2(Wickham 2016). Para realizar los modelos y obtener tablas con la informacion pertinente de cada modelo, asi como tambien para la seleccion de modelos se utilizaron los paquetes broom(Robinson 2017) y MuMln.

Resultados

La longitud total de lizard es 104.69

El peso total de lizard es 170.88

Graficos de diagrama de puntos

Table 1: Resumen del modelo lineal

term	estimate	std.error	statistic	p.value
(Intercept)	88.2539611	2.0688867	42.65771	0
Weight	0.0961706	0.0026769	35.92594	0

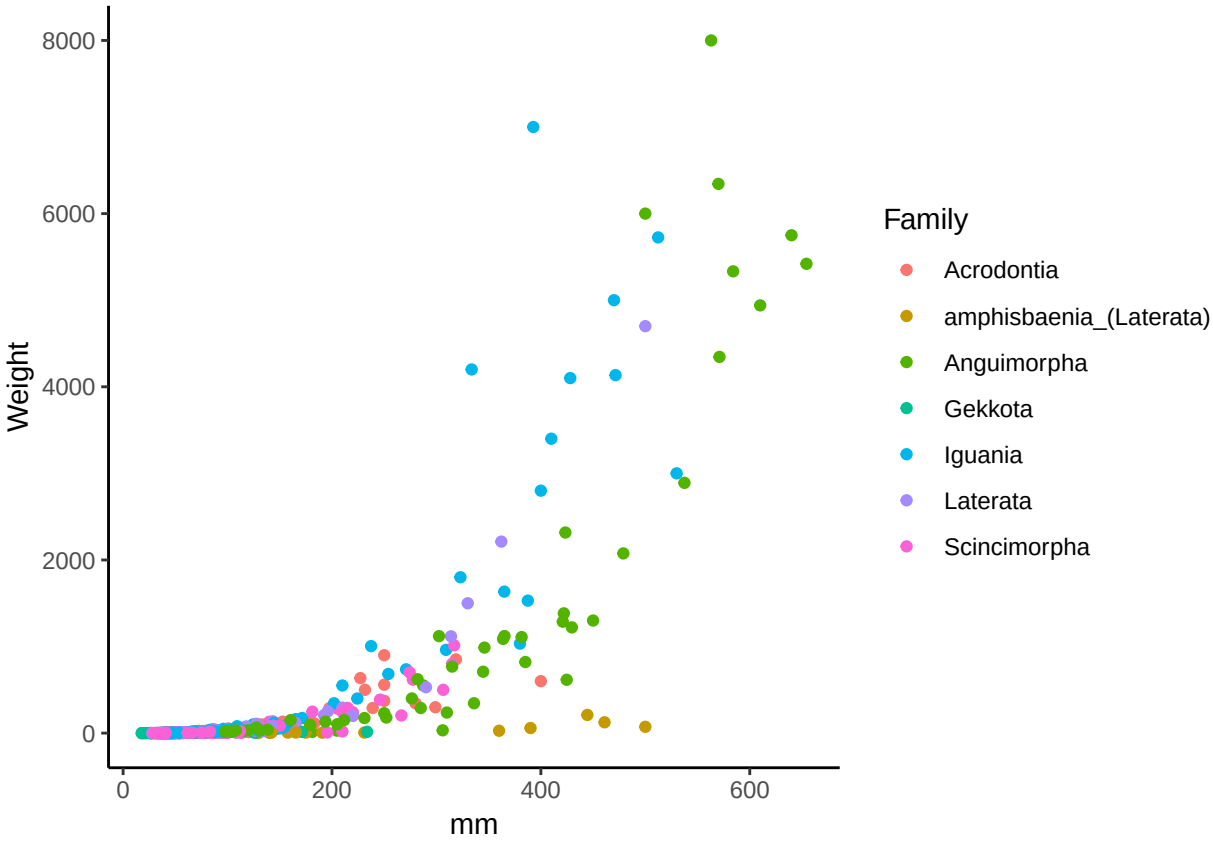


Figure 1: Modelo lineal lizard

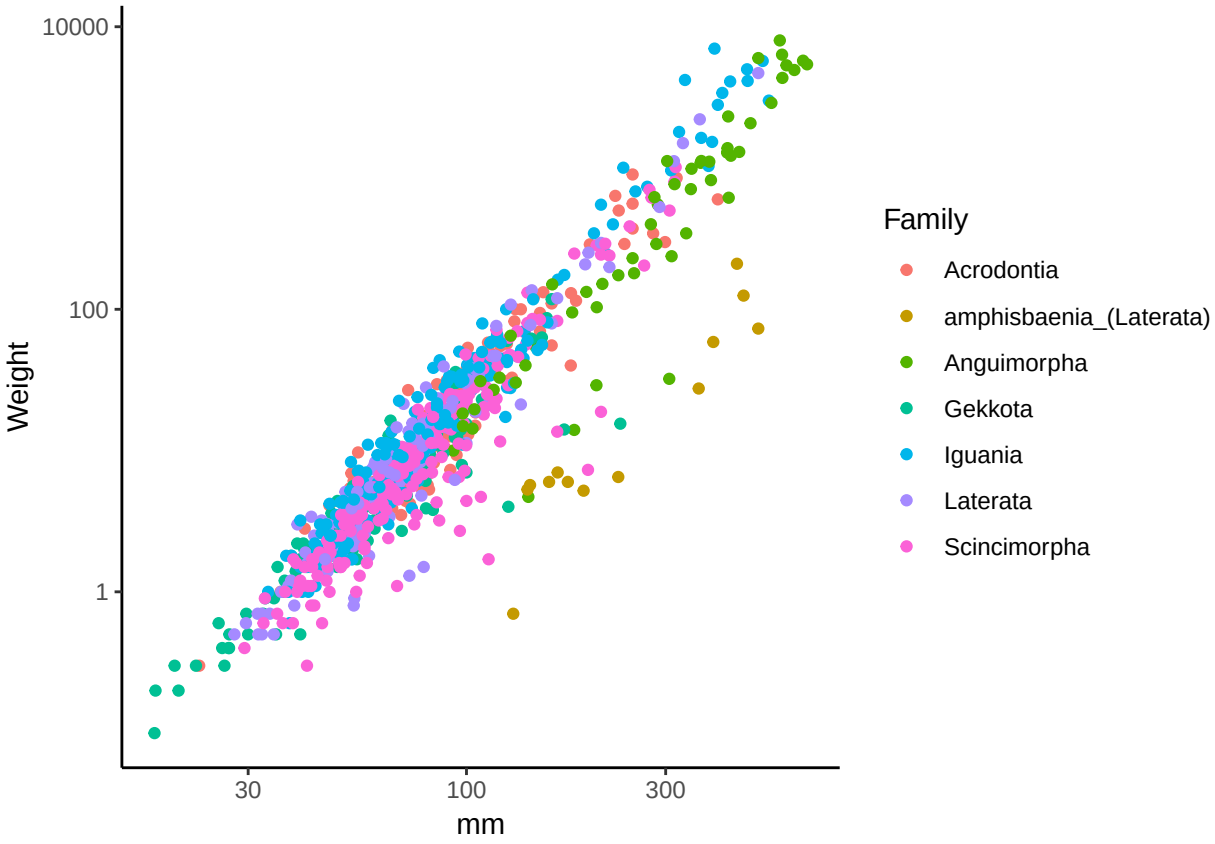
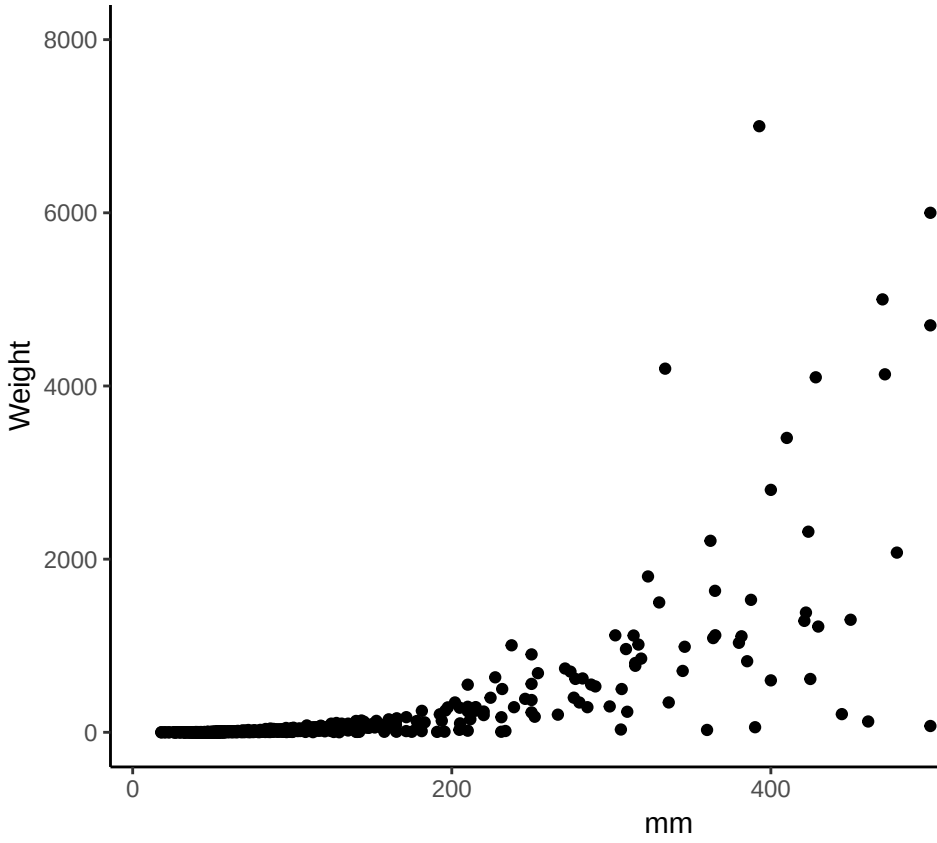


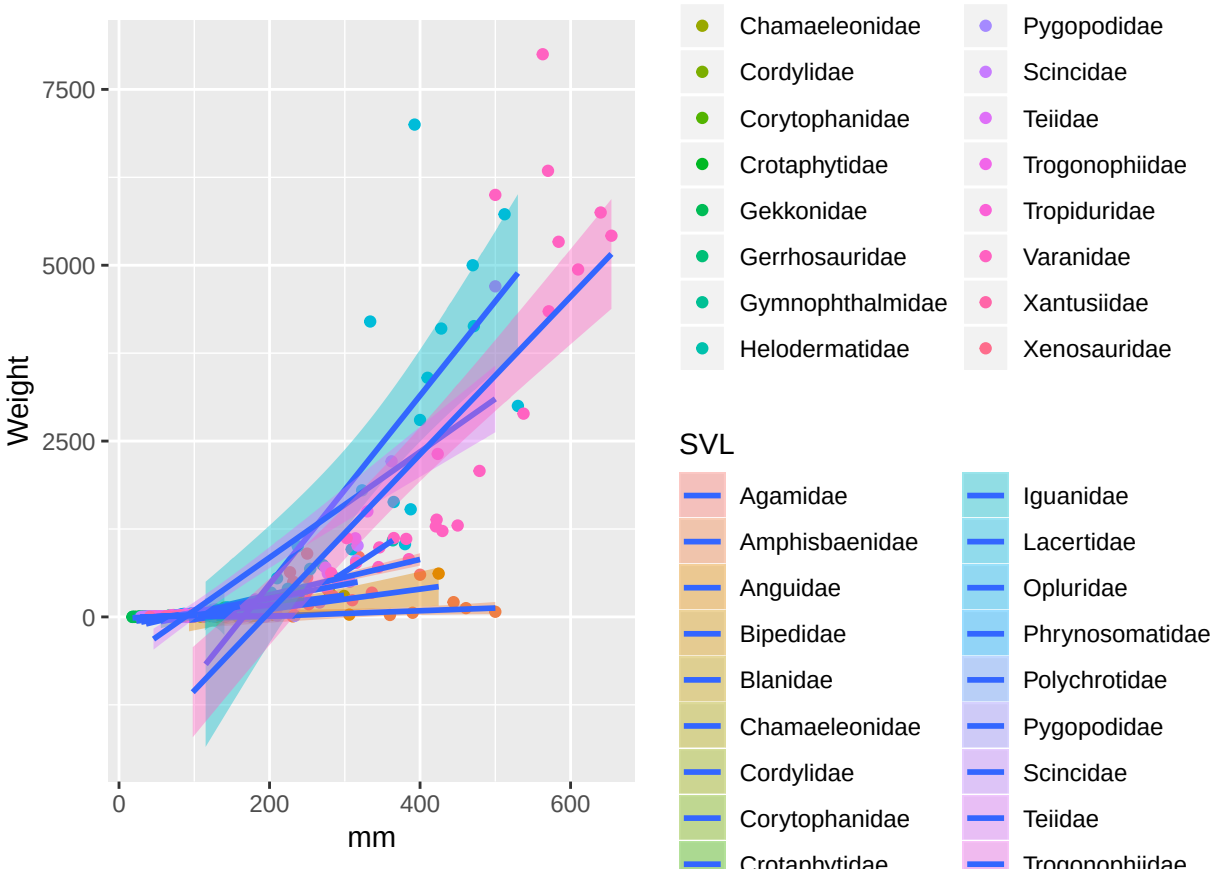
Figure 2: Modelo lineal lizard con escala log

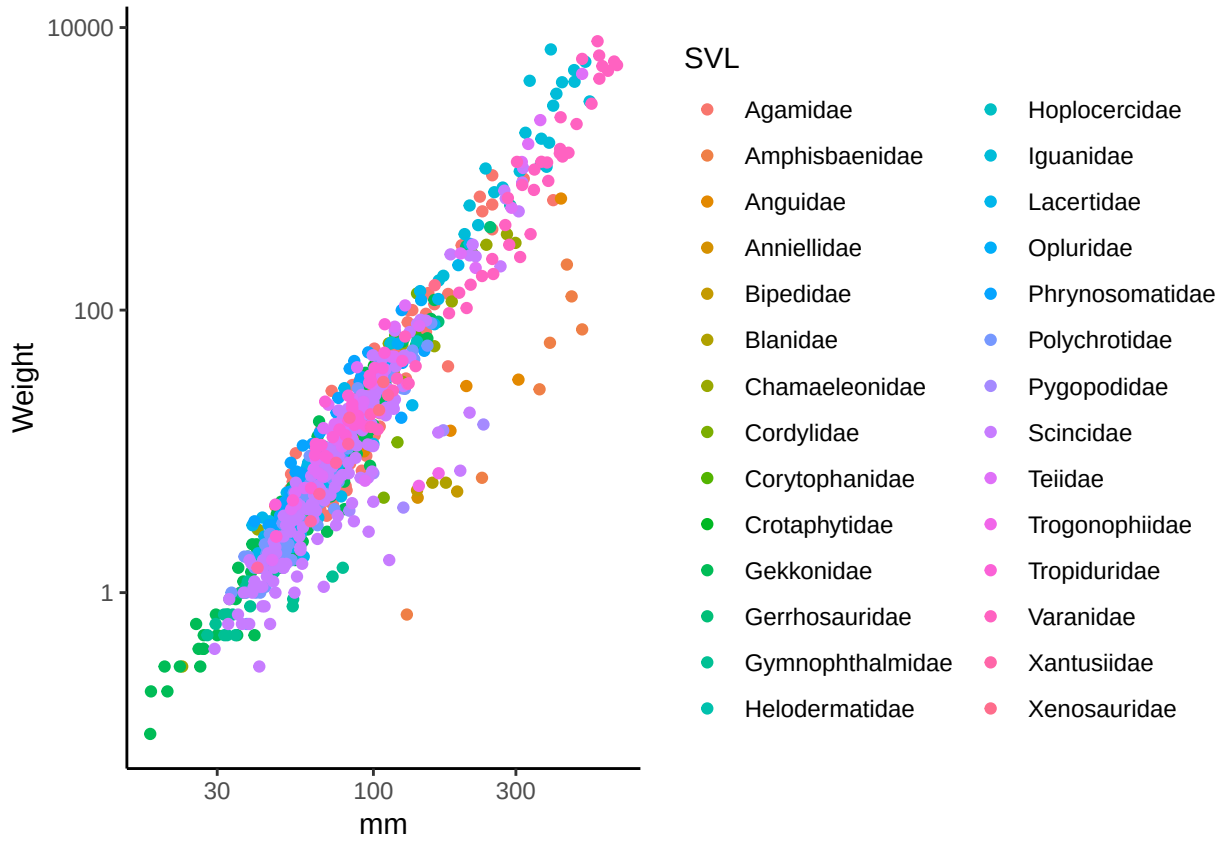
primero el grafico de puntos donde solo quiero mostrar la tendencia general de todas las categorias, realizando



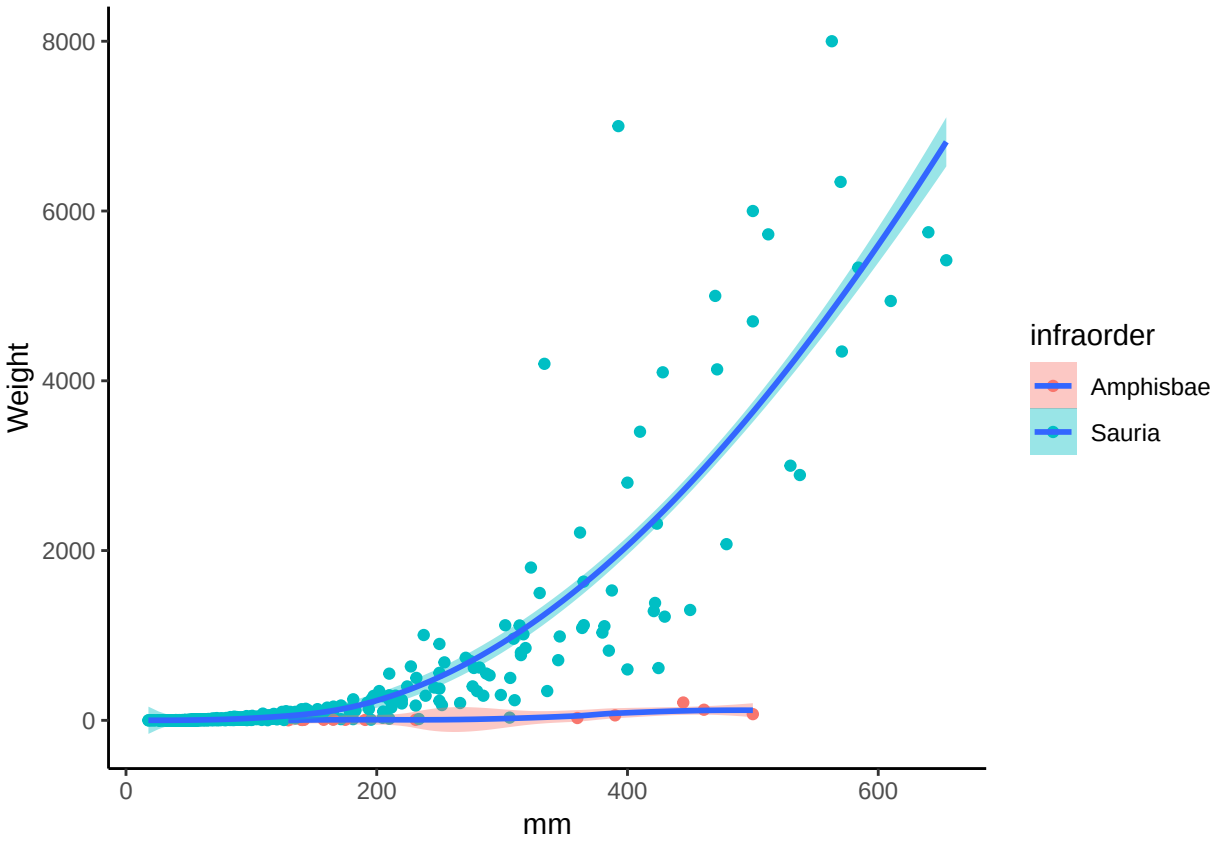
la relacion entre peso y longitud total (mm)

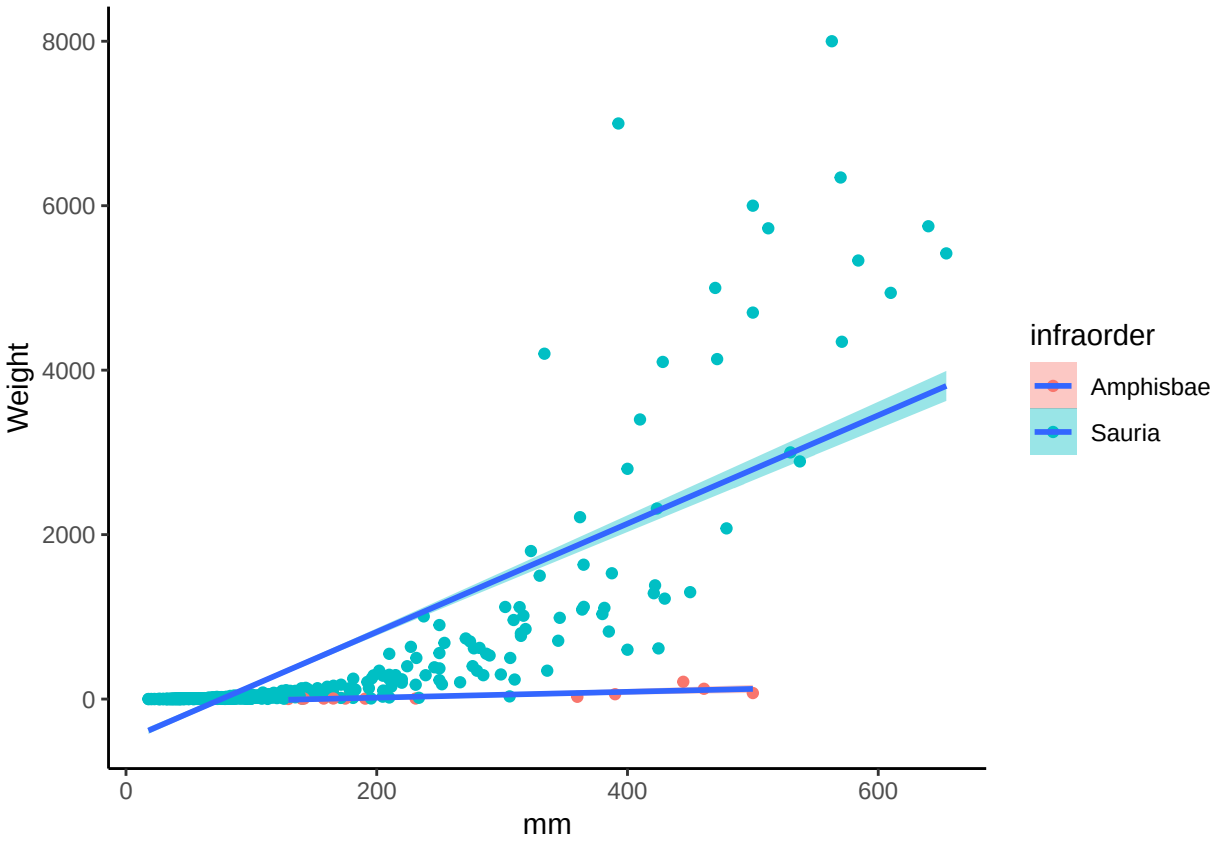
Luego si se quisiera ver la relacion del peso total y la longitud total en la categoria SVL, podria verse de dos maneras. donde queda claro que la segunda grafica es mucho mejor si se aplica log a ambas variables.

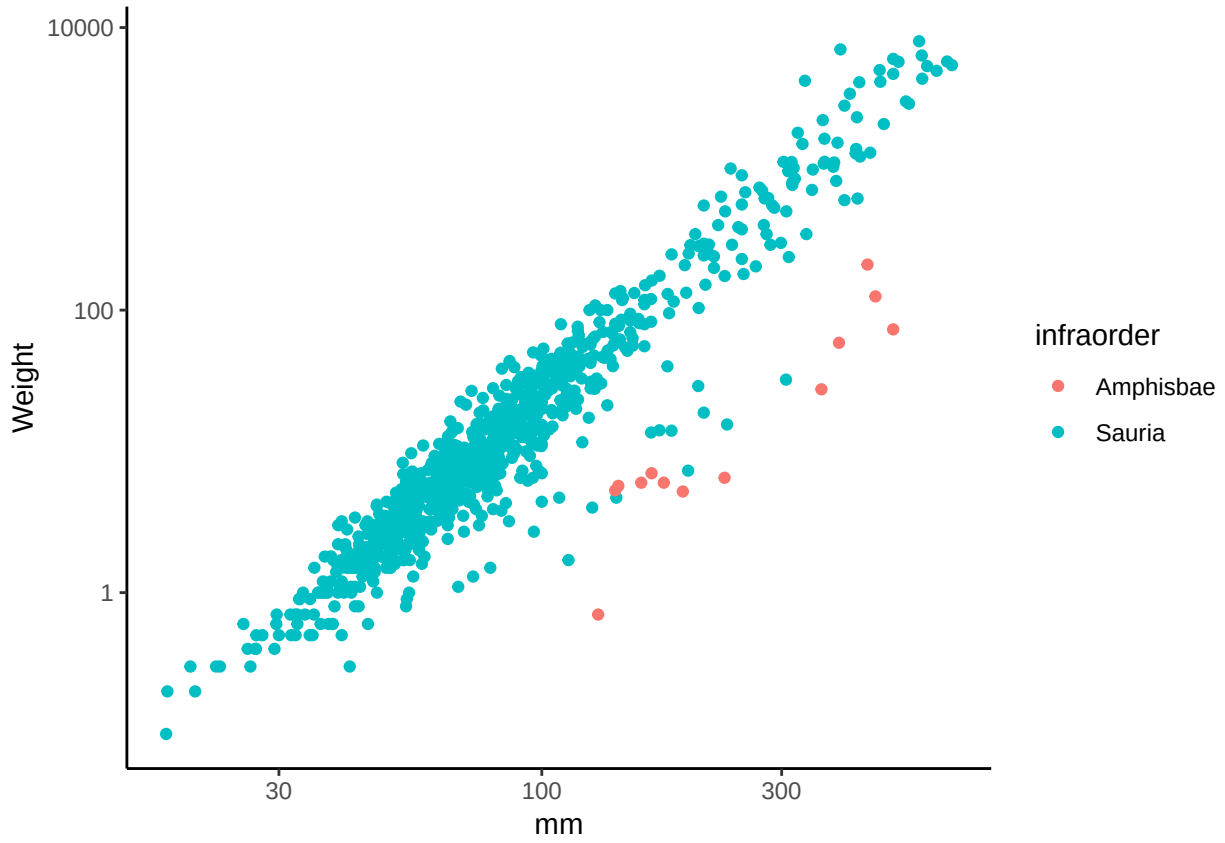

7500
5000
2500
0
Weight
0
200
400
600
mm
Chamaeleonidae
Cordylidae
Corytophanidae
Crotaphytidae
Gekkonidae
Gerrhosauridae
Gymnophthalmidae
Helodermatidae
Pygopodidae
Scincidae
Teiidae
Trogonophiidae
Tropiduridae
Varanidae
Xantusiidae
Xenosauridae
SVL
Agamidae
Amphisbaenidae
Anguidae
Bipedidae
Blanidae
Chamaeleonidae
Cordylidae
Corytophanidae
Crotaphytidae
Iguanidae
Lacertidae
Opluridae
Phrynosomatidae
Polychrotidae
Pygopodidae
Scincidae
Teiidae
Trogonophiidae

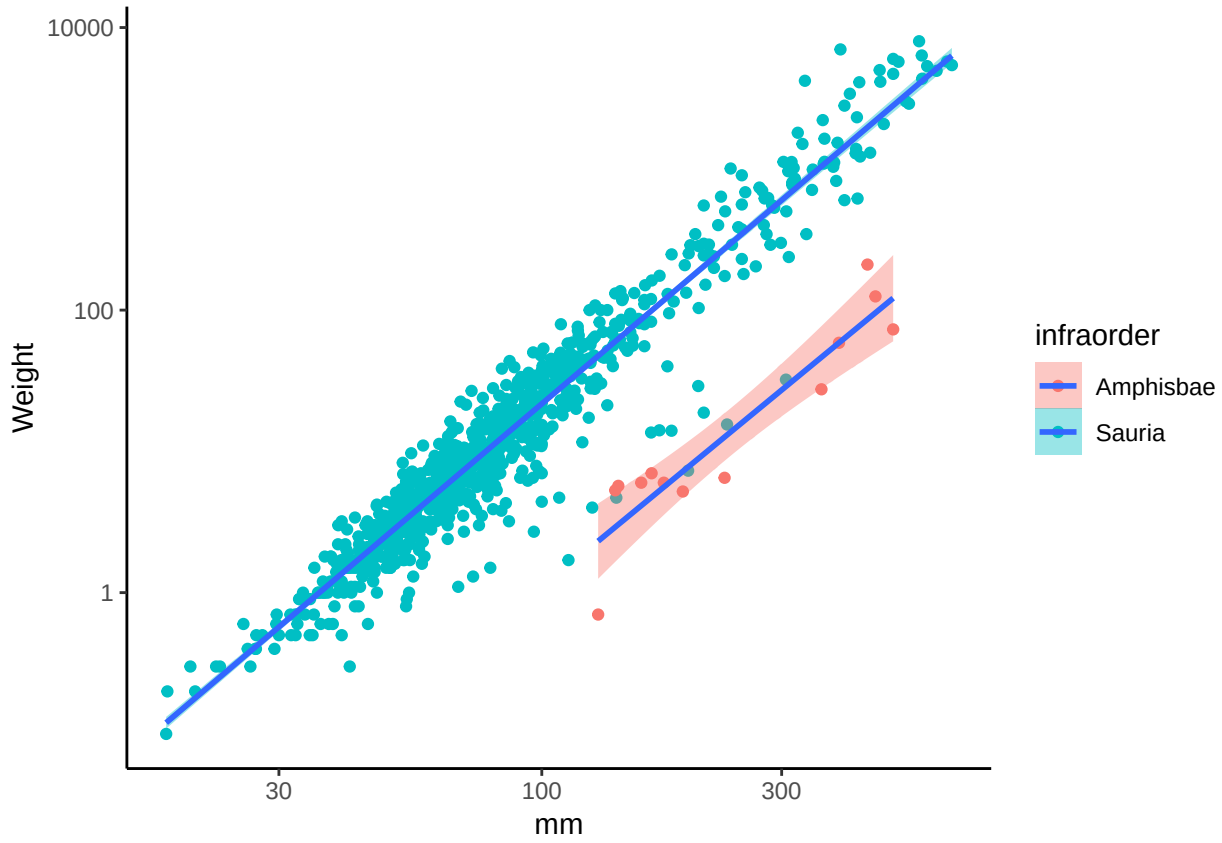


Ahora si quisieramos seleccionar una tendencia, podriamos utilizar algunos modelos lineales


8000
6000
4000
2000
0
Weight
0
200
400
600
mm
infraorder
Amphisbae
Sauria


8000
6000
4000
2000
0
Weight
0
200
400
600
mm
infraorder
Amphisbae
Sauria


10000
100
1
Weight
30
100
300
mm
infraorder
Amphisbae
Sauria



Modelos lm

Table 2: Resumen del modelo lineal de todas la variables utilizando mm y weigth

term	estimate	std.error	statistic	p.value
(Intercept)	88.2539611	2.0688867	42.65771	0
Weight	0.0961706	0.0026769	35.92594	0

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1     0.153         0.147  696.      27.3 5.34e-30     7 -7284. 14585.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 3: Resumen del modelo lineal de cada familia de lizard mm y weigth

term	estimate	std.error	statistic	p.value
(Intercept)	72.439474	65.23146	1.1104991	0.2670780
Familyamphisbaenia_(Laterata)	-31.162551	203.88594	-0.1528431	0.8785560
FamilyAnguimorpha	1197.378383	113.65474	10.5352265	0.0000000
FamilyGekkota	-63.151238	86.17217	-0.7328496	0.4638393

term	estimate	std.error	statistic	p.value
FamilyIguania	183.968622	81.02517	2.2705120	0.0234101
FamilyLaterata	6.142345	84.82370	0.0724131	0.9422891
FamilyScincimorpha	-32.143885	81.44334	-0.3946779	0.6931733

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1     0.357         0.353  76.2      84.1 9.18e-84     7 -5260. 10536.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 4: Resumen del modelo lineal de cada familia de lizard variable y= mm

term	estimate	std.error	statistic	p.value
(Intercept)	107.126403	7.139164	15.0054542	0.0000000
Familyamphisbaenia_(Laterata)	161.089750	22.314007	7.2192211	0.0000000
FamilyAnguimorpha	195.950382	12.438781	15.7531819	0.0000000
FamilyGekkota	-41.046077	9.430990	-4.3522552	0.0000150
FamilyIguania	-3.811784	8.867685	-0.4298512	0.6674059
FamilyLaterata	-27.099979	9.283409	-2.9191839	0.0035962
FamilyScincimorpha	-18.365913	8.913450	-2.0604718	0.0396379

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1     0.387         0.368  599.      20.7 4.10e-76    28 -7137. 14331.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 5: Resumen del modelo lineal de cada SVL de lizard variable y= weigth

term	estimate	std.error	statistic	p.value
(Intercept)	74.626374	62.84393	1.1874873	0.2353534
SVLAmphisbaenidae	-2.997802	235.14047	-0.0127490	0.9898309
SVLAnguidae	27.748626	221.07314	0.1255179	0.9001420
SVLAnniellidae	-69.926374	602.77784	-0.1160069	0.9076734
SVLBipedidae	-69.376374	428.53850	-0.1618906	0.8714288
SVLBlanidae	-68.626374	428.53850	-0.1601405	0.8728069
SVLChamaeleonidae	-10.839417	139.91100	-0.0774737	0.9382642
SVLCordylidae	-44.484707	184.11593	-0.2416125	0.8091363
SVLCorytophanidae	32.773626	306.26345	0.1070112	0.9148043
SVLCrotaphytidae	-33.951374	306.26345	-0.1108568	0.9117550
SVLGekkonidae	-65.324347	79.86040	-0.8179817	0.4135873
SVLGerrhosauridae	42.723626	221.07314	0.1932556	0.8468030
SVLGymnophthalmidae	-71.863517	119.23798	-0.6026898	0.5468691
SVLHelodermatidae	745.173626	428.53850	1.7388721	0.0824042
SVLHoplocercidae	-27.293040	351.77637	-0.0775863	0.9381746
SVLIguanidae	2118.169279	139.91100	15.1394044	0.0000000
SVLLacertidae	-57.551661	89.89049	-0.6402419	0.5221807

term	estimate	std.error	statistic	p.value
SVLOpluridae	-29.654945	235.14047	-0.1261159	0.8996688
SVLPhrynosomatidae	-55.166374	113.72842	-0.4850711	0.6277457
SVLPolychrotidae	-65.168642	87.48960	-0.7448730	0.4565458
SVLPygopodidae	-65.746374	275.36832	-0.2387579	0.8113484
SVLScincidae	-35.935978	77.32931	-0.4647136	0.6422507
SVLTeiidae	190.113161	110.93831	1.7136836	0.0869362
SVLTrogonophiidae	-68.276374	428.53850	-0.1593238	0.8734501
SVLTropiduridae	-53.798249	123.20857	-0.4366437	0.6624760
SVLVaranidae	1555.037912	111.83160	13.9051744	0.0000000
SVLXantusiidae	-64.397802	235.14047	-0.2738695	0.7842487
SVLXenosauridae	-7.759707	351.77637	-0.0220586	0.9824061

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic   p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>     <dbl> <int>  <dbl>  <dbl>
## 1     0.593         0.581  61.3      48.0 5.75e-153    28 -5051. 10159.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 6: Resumen del modelo lineal de cada SVL de lizard Y = mm

term	estimate	std.error	statistic	p.value
(Intercept)	107.126403	7.139164	15.0054542	0.0000000
Familyamphisbaenia_(Laterata)	161.089750	22.314007	7.2192211	0.0000000
FamilyAnguimorpha	195.950382	12.438781	15.7531819	0.0000000
FamilyGekkota	-41.046077	9.430990	-4.3522552	0.0000150
FamilyIguania	-3.811784	8.867685	-0.4298512	0.6674059
FamilyLaterata	-27.099979	9.283409	-2.9191839	0.0035962
FamilyScincimorpha	-18.365913	8.913450	-2.0604718	0.0396379

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl>  <dbl>
## 1  0.000426     -0.000669  754.     0.389   0.533     2 -7360. 14726.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 7: Resumen del modelo lineal de cada infraorden de lizard weigth

term	estimate	std.error	statistic	p.value
(Intercept)	41.27692	209.2327	0.1972776	0.8436542
infraorderSauria	131.47363	210.7350	0.6238812	0.5328615

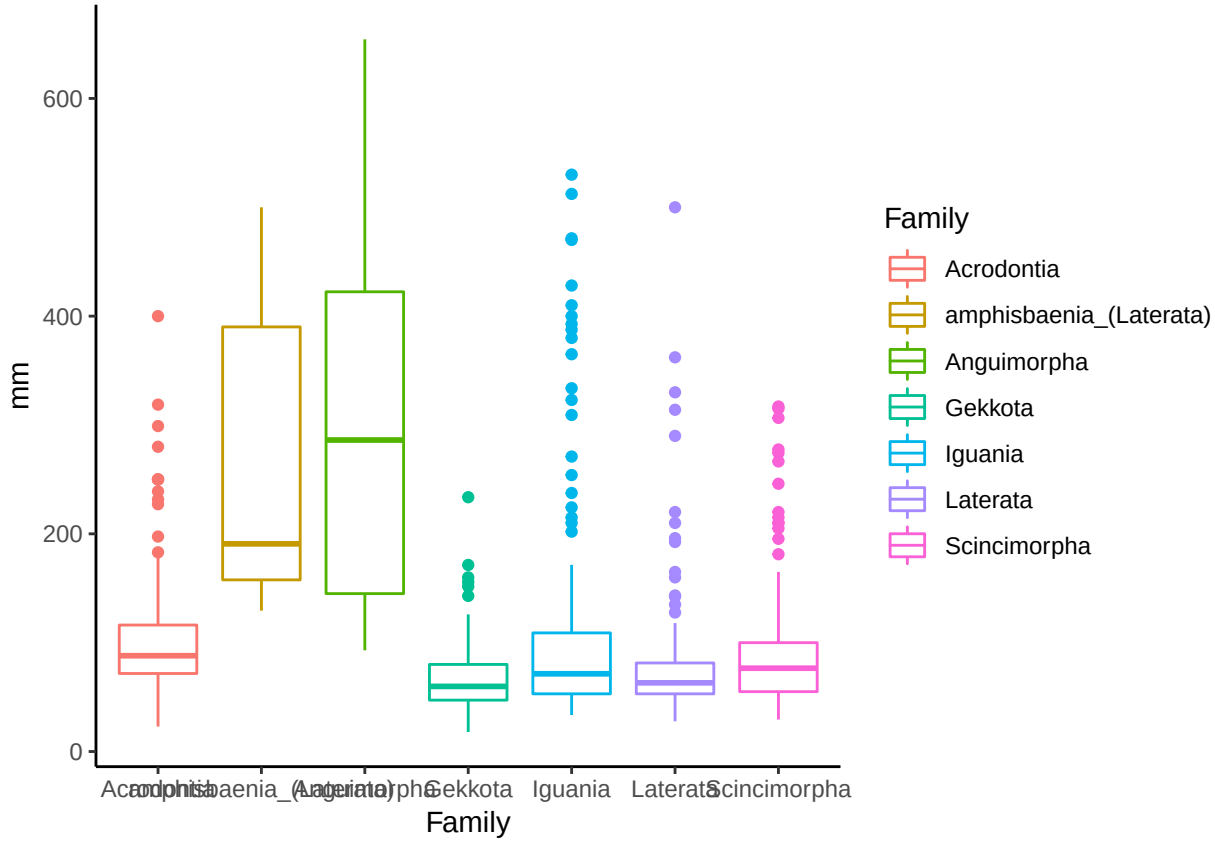
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1    0.0430        0.0419  92.8      41.0 2.45e-10     2 -5442. 10891.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 8: Resumen del modelo lineal de cada infraorder de lizard mm

term	estimate	std.error	statistic	p.value
(Intercept)	41.27692	209.2327	0.1972776	0.8436542
infraorderSauria	131.47363	210.7350	0.6238812	0.5328615

Graficos Boxplot

Para el caso de la variable Familia, al tener tantos datos extremos el hecho de cambiar la escala mejora la visualizacion de datos



Aqui tenemos un boxplot aplicando los datos normales y luego transformando los ejes a log. la diferencia es

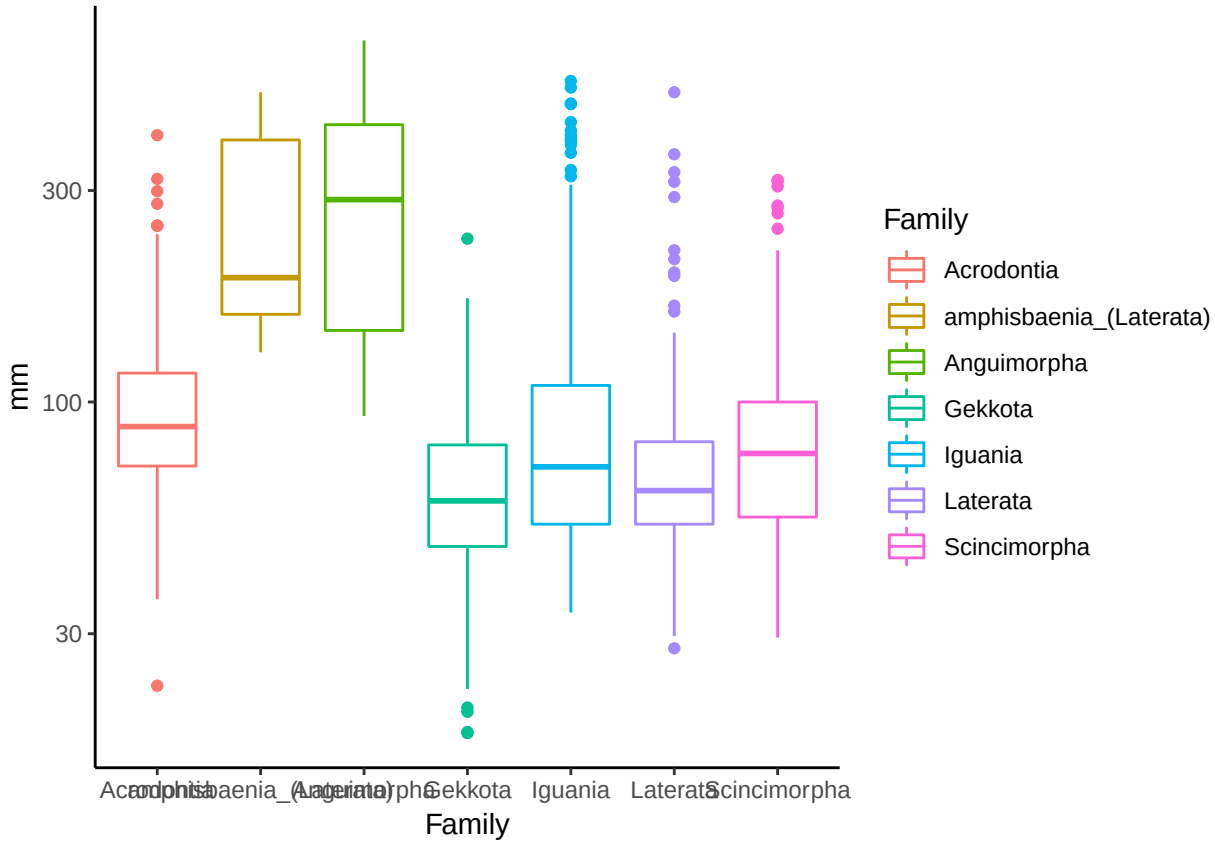
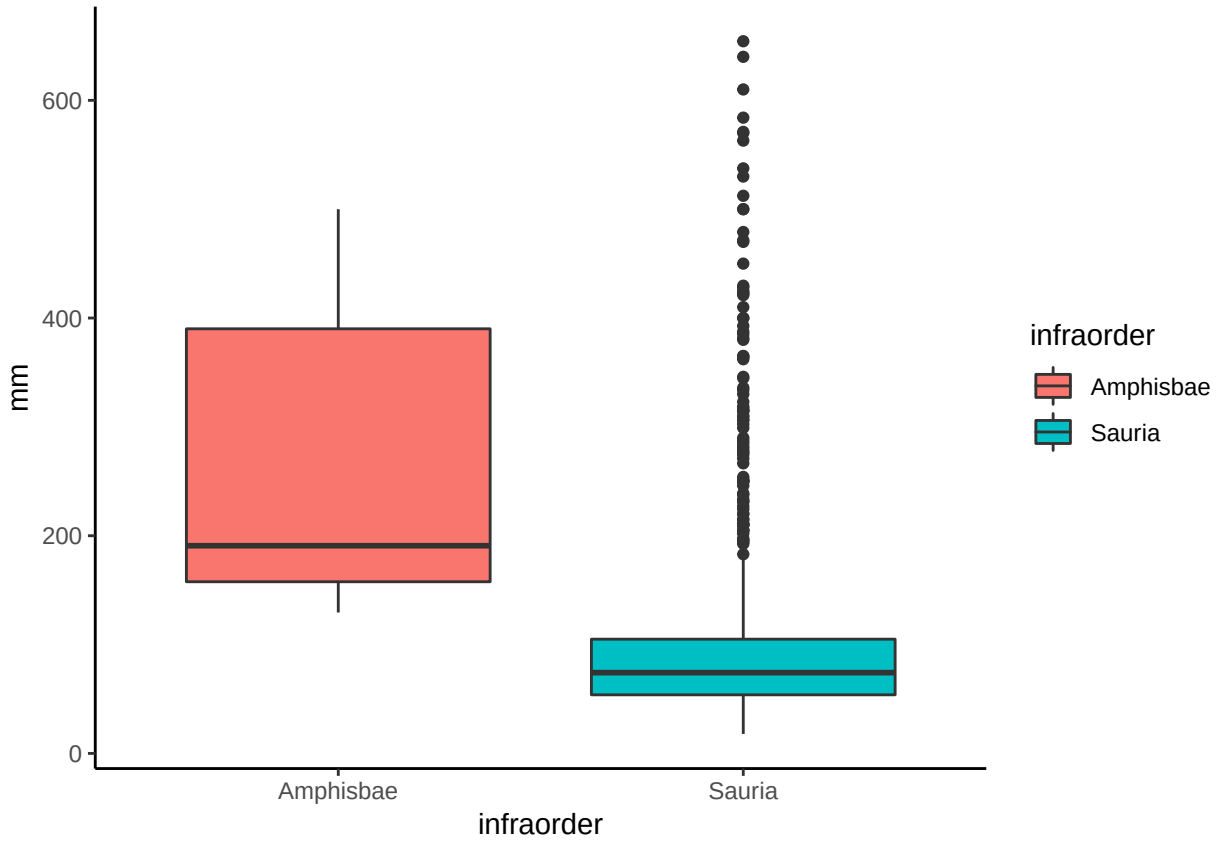


Figure 3: Modelo lineal lizard en escala log



clara

Estos resultados cambiando la amplitud de los ejes se vuelven a repetir para la variable limbs

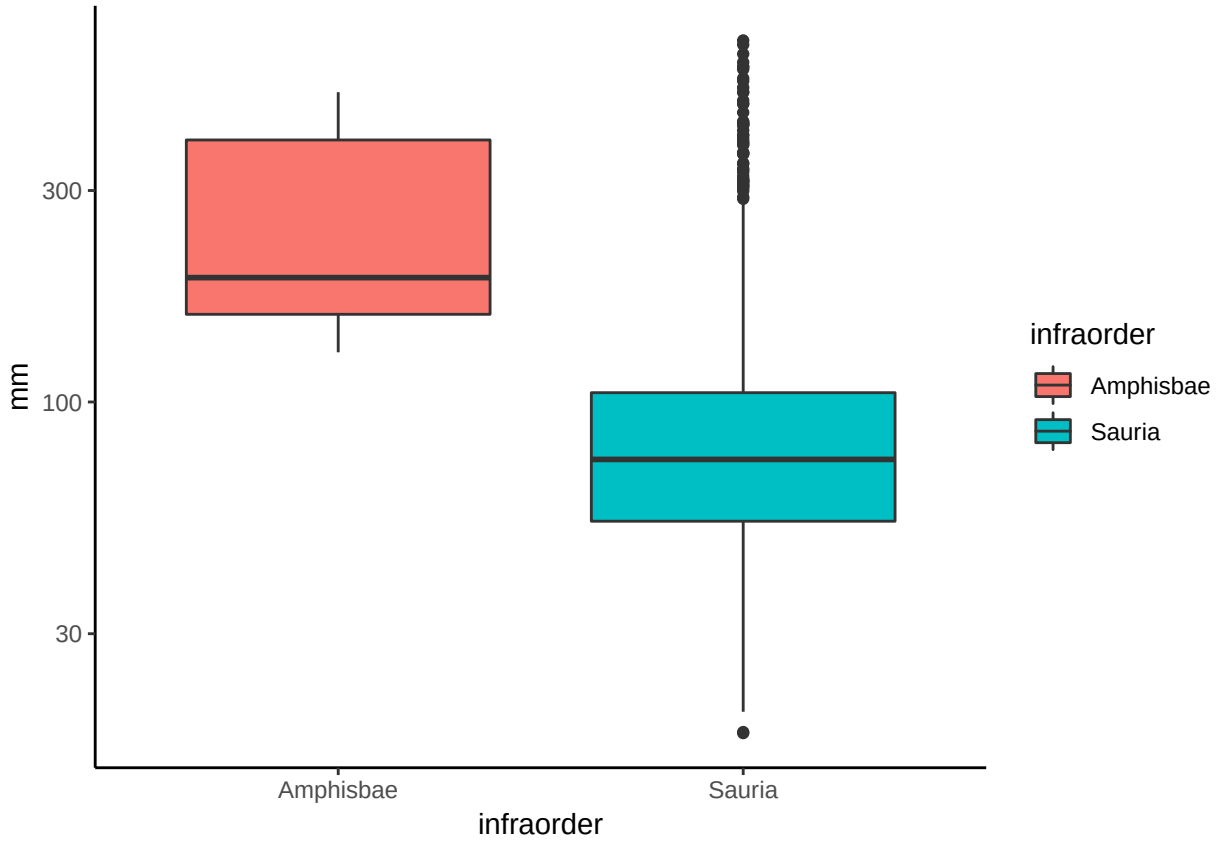
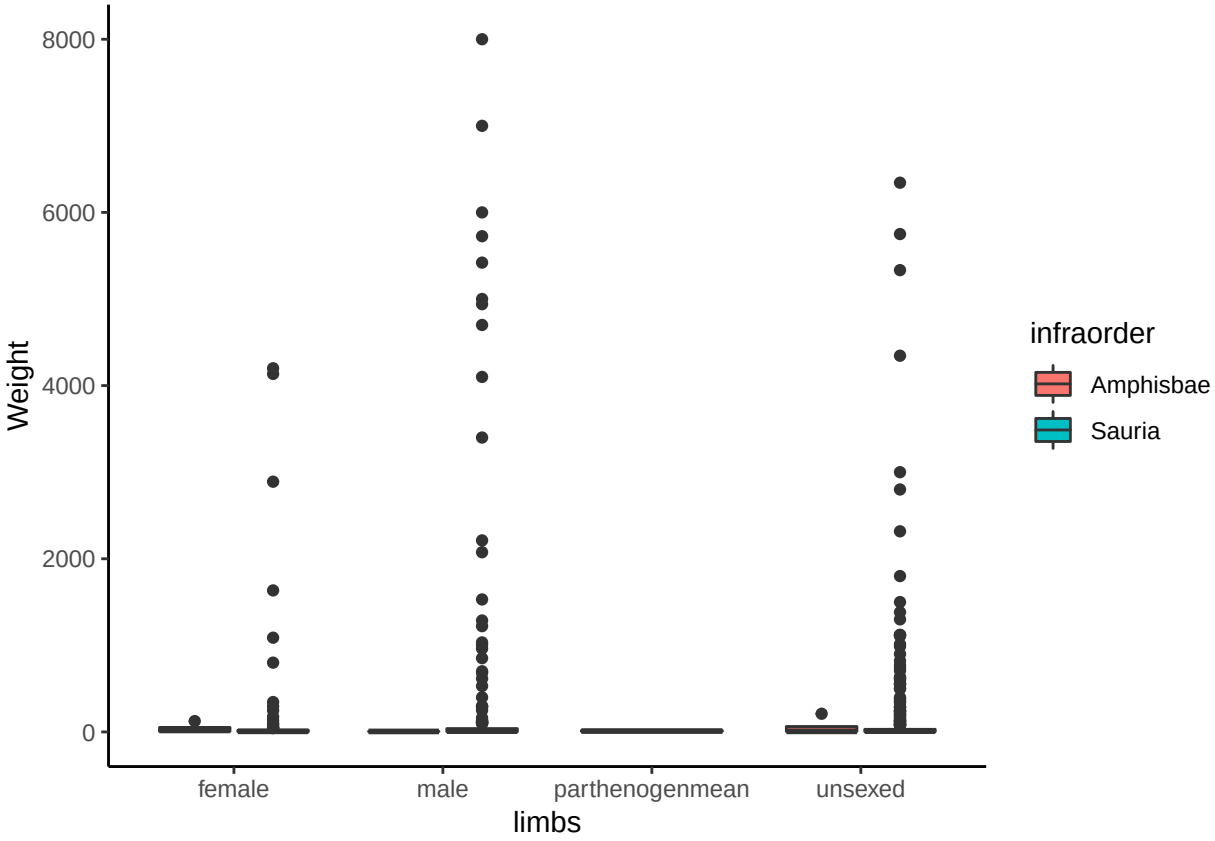


Figure 4: Modelo lineal lizard en escala log



Modelos aov y t.test

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl>  <dbl>
## 1  0.000426     -0.000669  754.     0.389   0.533     2 -7360. 14726.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 9: Resumen del modelo ANDEVA de cada infraorden de lizard weigth

term	df	sumsq	meansq	statistic	p.value
infraorder	1	221516.5	221516.5	0.3892277	0.5328615
Residuals	913	519604733.6	569118.0	NA	NA

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1    0.0430        0.0419  92.8      41.0 2.45e-10     2 -5442. 10891.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

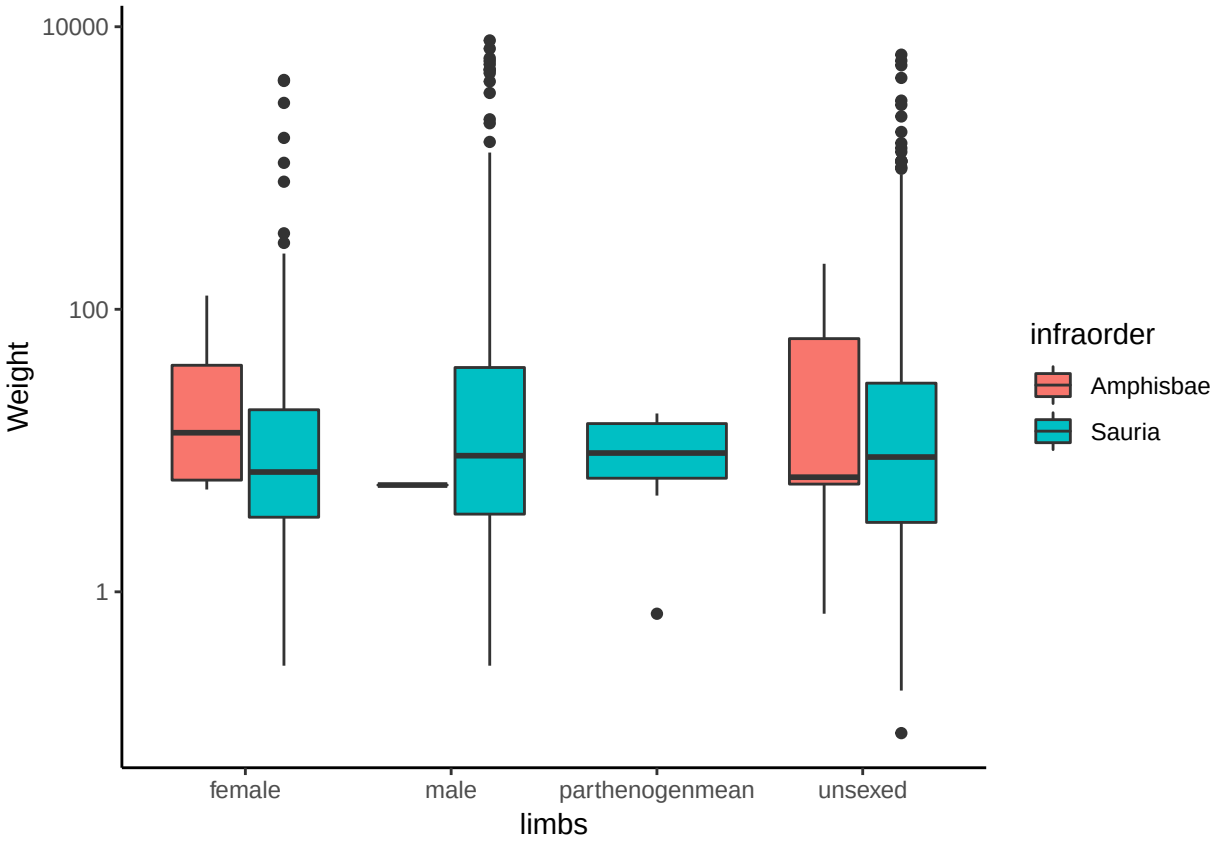


Figure 5: Modelo lineal lizard en escala log

Table 10: Resumen del modelo ANDEVA de cada infraorden de lizard mm

term	df	sumsq	meansq	statistic	p.value
infraorder	1	352649.9	352649.920	40.98367	0
Residuals	913	7856040.6	8604.645	NA	NA

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1    -131.      41.3      173.     -4.28 4.00e-5      109.    -192.
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

Table 11: Resumen del modelo T.TEST de cada infraorden de lizard WEIGTH

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	a
-131.4736	41.27692	172.7506	-4.281657	4e-05	109.3017	-192.3305	-70.61672	Welch Two Sample t-test	t

```
## # A tibble: 1 x 10
##   estimate estimate1 estimate2 statistic p.value parameter conf.low
##      <dbl>     <dbl>     <dbl>     <dbl>   <dbl>     <dbl>    <dbl>
## 1     166.      268.      102.      4.25 0.00109      12.1     81.0
## # ... with 3 more variables: conf.high <dbl>, method <chr>,
## #   alternative <chr>
```

Table 12: Resumen del modelo T.TEST de cada infraorden de lizard mm

estimate	estimate1	estimate2	statistic	p.value	parameter	conf.low	conf.high	method	a
165.8851	268.2162	102.331	4.253986	0.0010894	12.14936	81.03755	250.7327	Welch Two Sample t-test	t

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>   <dbl> <int>  <dbl>  <dbl>
## 1  0.000426     -0.000669  754.     0.389   0.533     2 -7360. 14726.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 13: Resumen del modelo ANDEVAde cada infraorden de lizard mm

term	df	sumsq	meansq	statistic	p.value
infraorder	1	221516.5	221516.5	0.3892277	0.5328615
Residuals	913	519604733.6	569118.0	NA	NA

```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic  p.value    df logLik    AIC
```

```
## *     <dbl>         <dbl> <dbl>     <dbl>    <dbl> <int>  <dbl>  <dbl>
## 1    0.0430        0.0419  92.8      41.0 2.45e-10     2 -5442. 10891.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

Table 14: Resumen del modelo ANDEVAde cada infraorden de lizard mm

term	df	sumsq	meansq	statistic	p.value
infraorder	1	352649.9	352649.920	40.98367	0
Residuals	913	7856040.6	8604.645	NA	NA

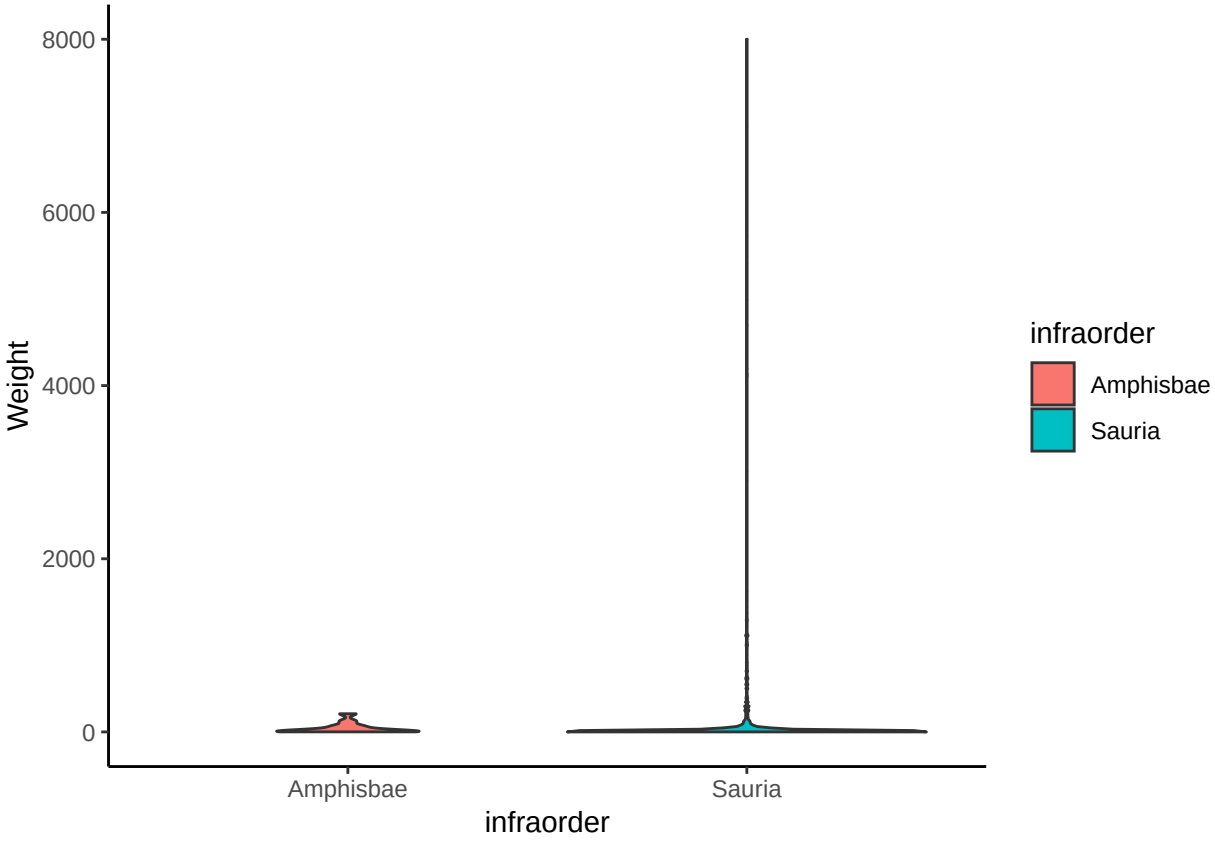
```
## # A tibble: 1 x 11
##   r.squared adj.r.squared sigma statistic   p.value    df logLik    AIC
## *     <dbl>         <dbl> <dbl>     <dbl>     <dbl> <int>  <dbl>  <dbl>
## 1     0.639         0.638  57.0      806. 2.44e-202     3 -4997. 10001.
## # ... with 3 more variables: BIC <dbl>, deviance <dbl>, df.residual <int>
```

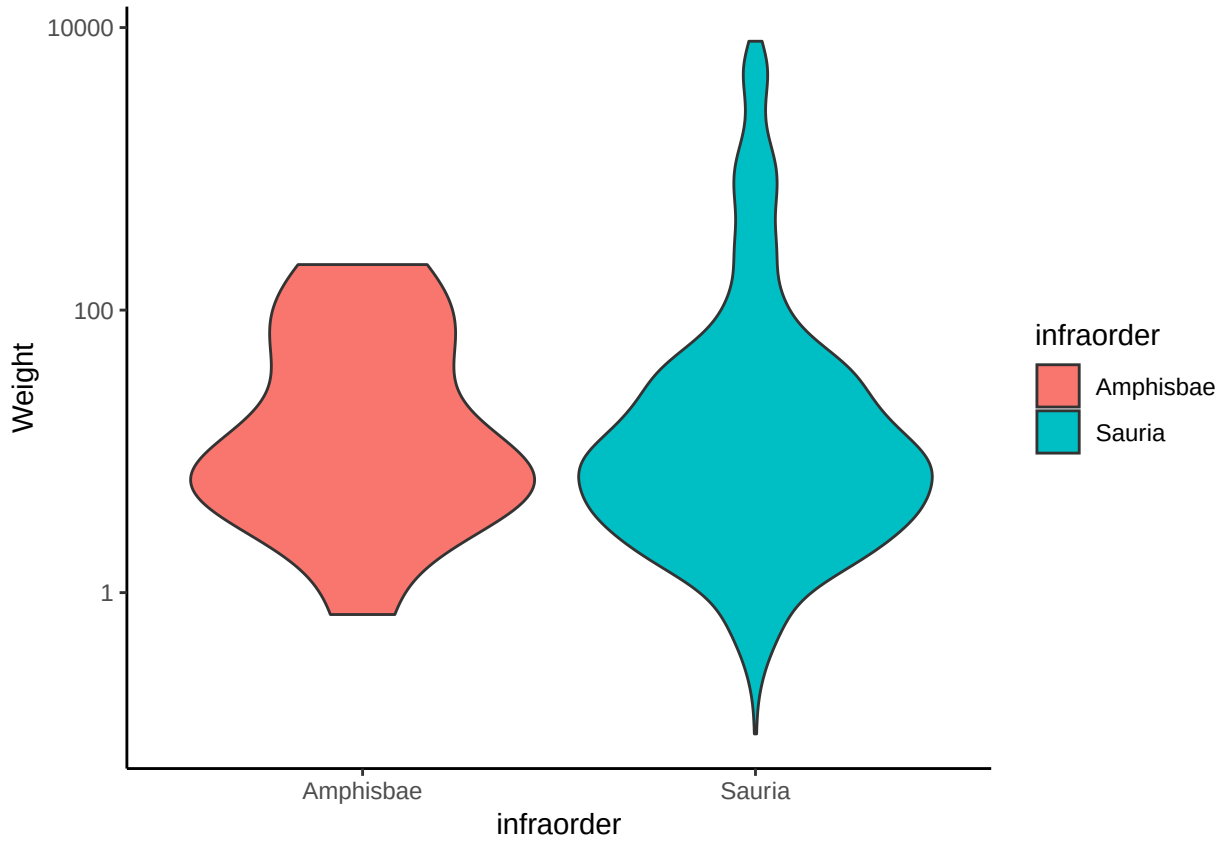
Table 15: Resumen del modelo ANDEVAde cada infraorden de lizard mm

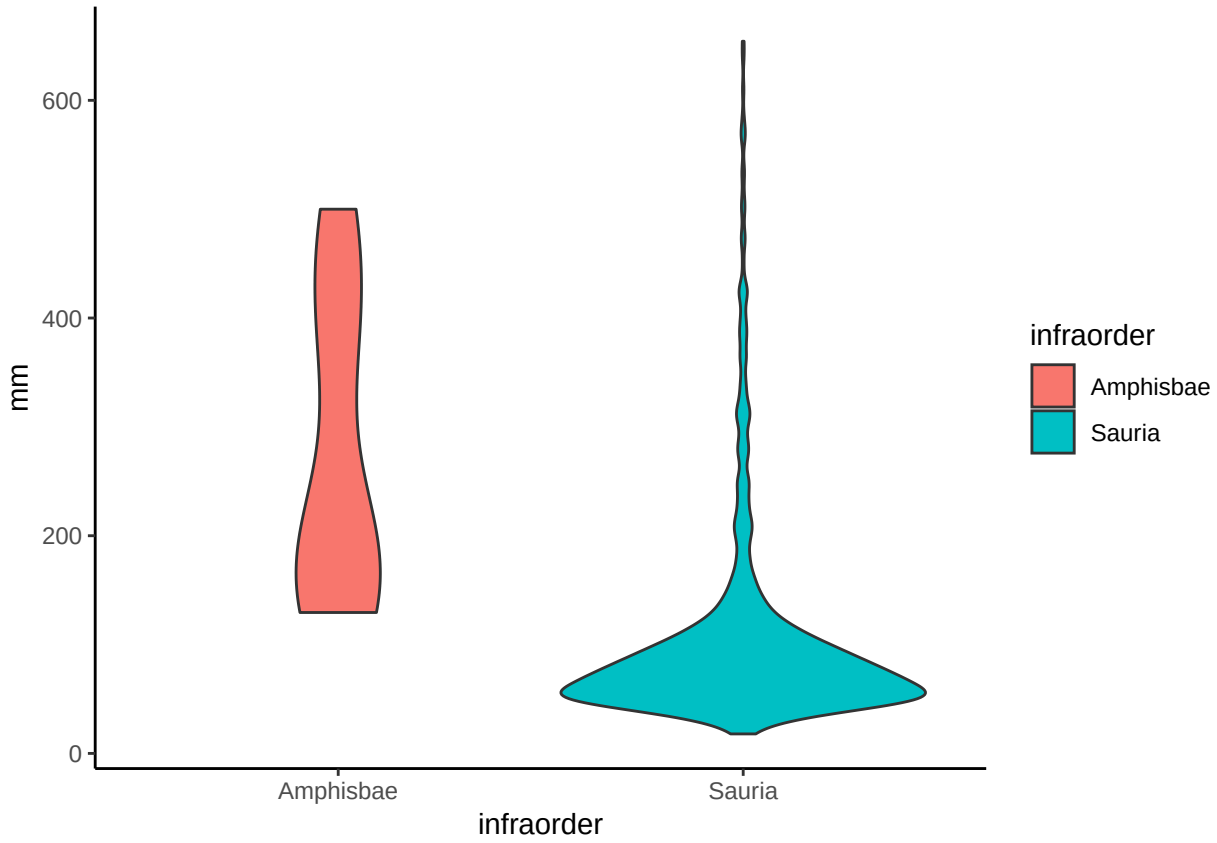
term	df	sumsq	meansq	statistic	p.value
infraorder:Weight	2	5242901	2621450.363	806.1133	0
Residuals	912	2965790	3251.962	NA	NA

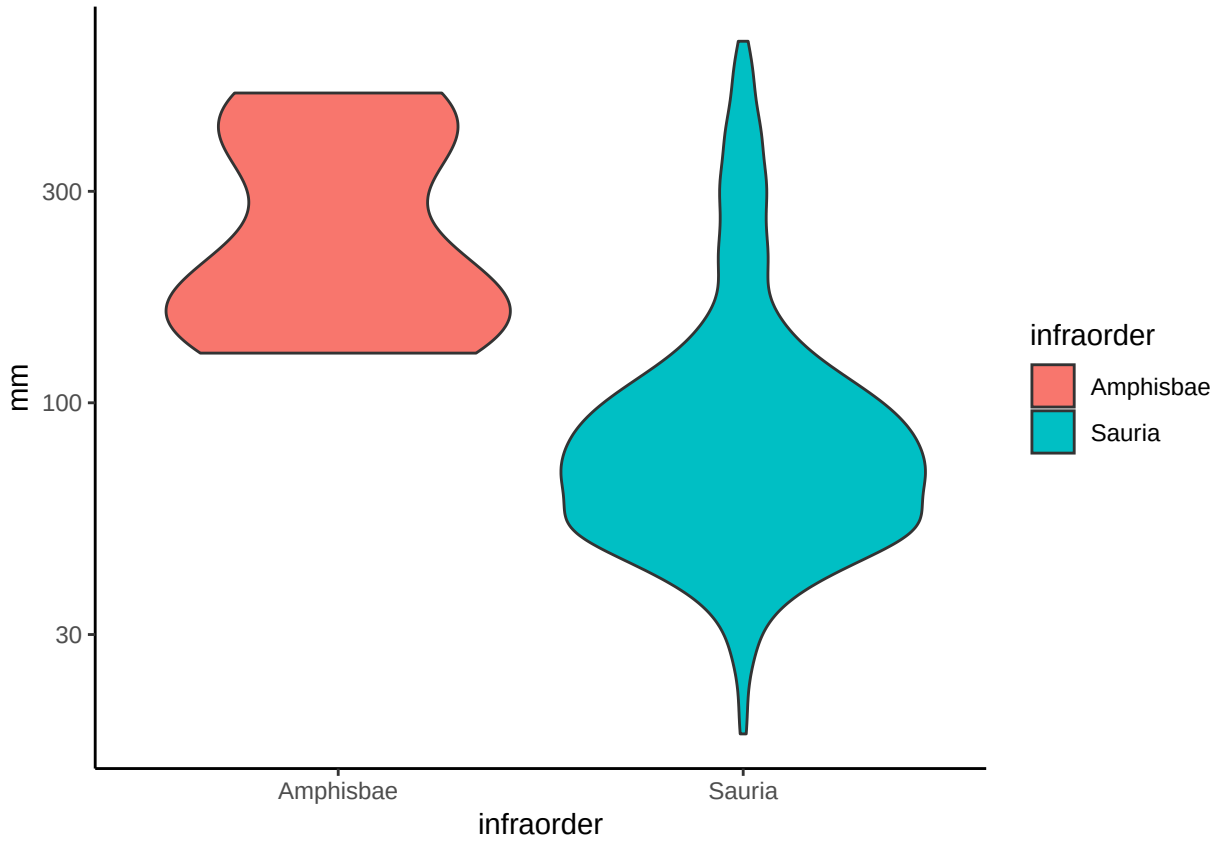
Graficos de Violin

En los graficos de violin tambien es apreciable la difrencia de escala en los distintos graficos. ESte tipo de graficos muestran donde se concentran la mayoria de los datos y tiene la capacidad de mostrar como estan distribuidos para una msiam categoria similar a una curva normal.


8000
6000
4000
2000
0
Weight
Amphisbae
Sauria
infraorder
infraorder
Amphisbae
Sauria


10000
100
1
Weight
Amphisbae
Sauria
infraorder
infraorder
Amphisbae
Sauria


600
400
200
0
mm
Amphisbae
Sauria
infraorder
infraorder
Amphisbae
Sauria



Conclusion

la eleccion de mostrar los datos va adpeender del tipo de variable que tengamos, estas pueden ser de razon, categoricas principalmente. Junto con esto, no siempre se tienen los datos como uno los quisiera por lo que es necesario realizar un exploracion previa.

Referencias

Allaire, J, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. n.d. "Rmarkdown: Dynamic Documents for R; 2018." *URL Https://CRAN. R-Project. Org/Package= Rmarkdown. R Package Version* 1.

Berke, Olaf. 2006. "A Primer of Ecological Statistics by Hj Gotelli and Am Ellison." *Biometrics* 62 (1). Wiley Online Library: 308–8.

Hlavac, Marek. 2018. "Stargazer: Well-Formatted Regression and Summary Statistics Tables. Bratislava, Slovakia: Central European Labour Studies Institute (Celsi)."

Meiri, S. 2010. "Length–weight Allometries in Lizards." *Journal of Zoology* 281 (3). Wiley Online Library: 218–26.

Robinson, David. 2017. "Broom: Convert Statistical Analysis Objects into Tidy Data Frames." *R Package*

Version 0.4 2.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer.

———. 2017. "Tidyverse: Easily Install and Load'tidyverse'packages." *R Package Version* 1 (1).

Wickham, Hadley, and others. 2014. "Tidy Data." *Journal of Statistical Software* 59 (10). Foundation for Open Access Statistics: 1–23.

Xie, Yihui. 2015. "Dynamic Documents with R and Knitr. Boca Raton, Florida: Chapman; Hall/Crc."

Zar, Jerrold H, and others. 1999. *Biostatistical Analysis*. Pearson Education India.