




# Proyecto de análisis de datos

La CoNGA Physics

Juan David Hernández

# Obtención de datos


Se extrajeron los datos de la siguiente fuente <https://gitmilab.redclara.net/tutoriales/tutorial-mkdocs/-/tree/master/docs/modulo-datos/claseMD07/materialesMD07/data>, para el análisis se utilizaron los archivos species.csv y surveys.csv.









added notebooks to MD05

Juan Carlos Basto Pineda authored 2 weeks ago

2ce5993c



Name	Last commit	Last update
..		
 analisis.csv	added notebooks to MD05	2 weeks ago
 brics.csv	added notebooks to MD05	2 weeks ago
 cars.csv	added notebooks to MD05	2 weeks ago
 census.csv	added notebooks to MD05	2 weeks ago
 species.csv	added notebooks to MD05	2 weeks ago
 surveys.csv	added notebooks to MD05	2 weeks ago

# Transformación de los datos

En el repositorio <https://gitmilab.redclara.net/hernandezj/ejercicios-clase-08-datos.git> se encuentra el código (escrito en python3 en un jupyternotwbook), que contiene el proceso de transformación y análisis de los datos.

Se leyeron los datos de formato csv utilizando la librería pandas y se analizaron como dataframes de pandas.

```
df_species.head()
```

	species_id	genus	species	taxa
0	AB	Amphispiza	bilineata	Bird
1	AH	Ammospermophilus	harrisi	Rodent
2	AS	Ammodramus	savannarum	Bird
3	BA	Baiomys	taylori	Rodent
4	CB	Campylorhynchus	brunneicapillus	Bird

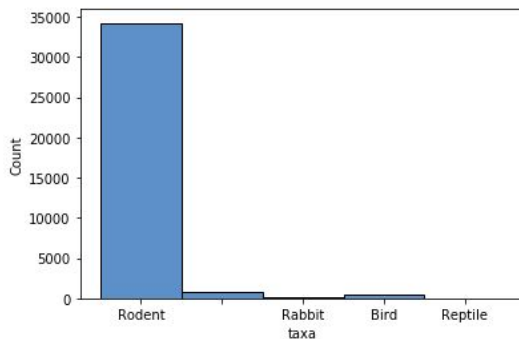
```
df_surveys.head()
```

	record_id	month	day	year	plot_id	species_id	sex	hindfoot_length	weight
0	1	7	16	1977	2	NL	M	32.0	NaN
1	2	7	16	1977	3	NL	M	33.0	NaN
2	3	7	16	1977	2	DM	F	37.0	NaN
3	4	7	16	1977	7	DM	M	36.0	NaN
4	5	7	16	1977	3	DM	M	35.0	NaN

Observando los nombres de las columnas y los tipos de datos, nos damos cuenta que se trata del registro de diferentes animales de varias especies y taxones. De los cuales se anotó su sexo, peso y longitud de su retropie, además se registró la fecha en la cual se hizo el análisis de cada dato.

```
sns.histplot(data=df_surveysspecie,x=df_surveysspecie['taxa'])
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7ff4ff3458d0>
```



```
df_taxa_peso_nan=df_surveysspecie[df_surveysspecie['weight']>0].groupby('taxa').count()  
df_taxa_pie_nan=df_surveysspecie[df_surveysspecie['hindfoot_length']>0].groupby('taxa').count()  
df_taxa_sexo_nan=df_surveysspecie[df_surveysspecie['hindfoot_length']>0].groupby('taxa').count()
```

```
df_taxa_peso_nan.shape
```

```
(1, 11)
```

```
df_taxa_pie_nan.shape
```

```
(1, 11)
```

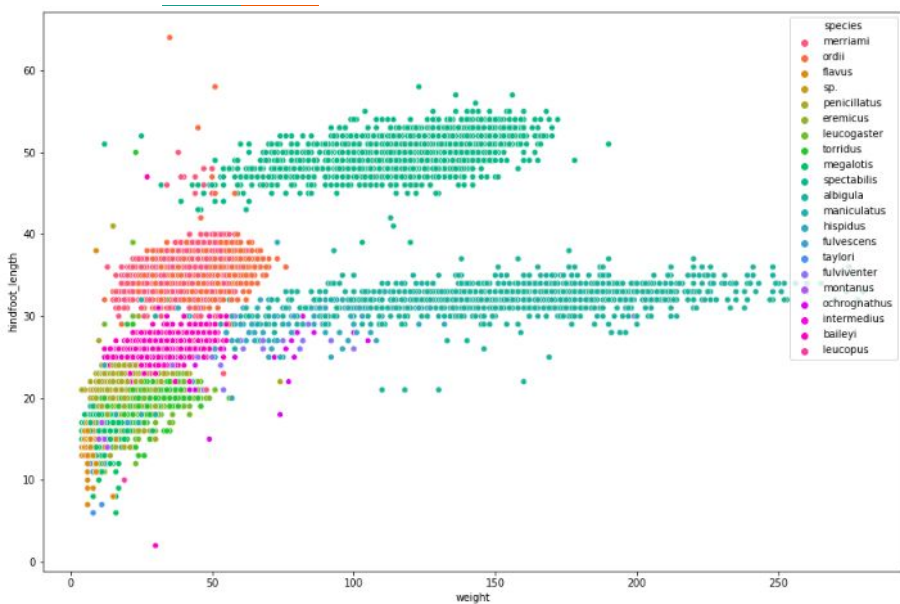
```
df_taxa_sexo_nan.shape
```

```
(1, 11)
```

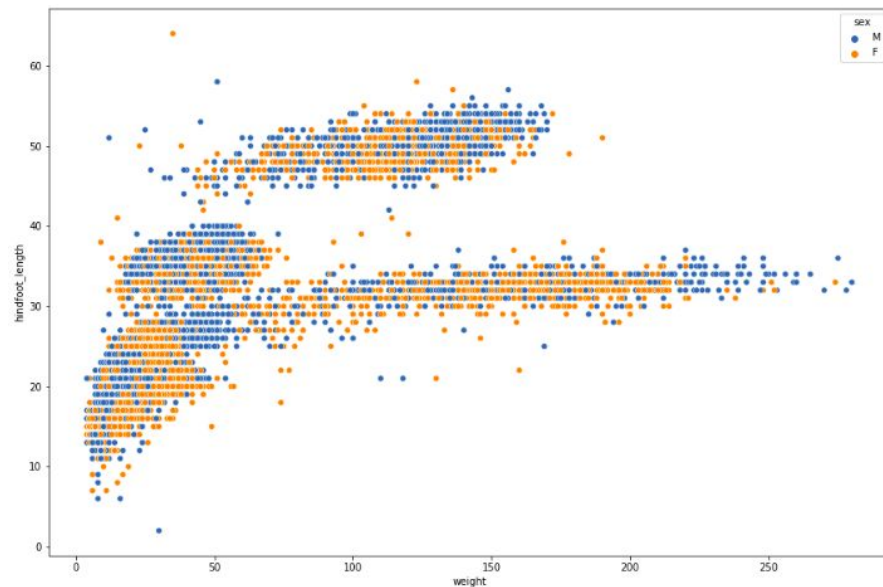
Un primer análisis nos informa que la mayor parte de los individuos son roedores y, además, solo se tiene información cuantitativa de este taxón. Debido a esto, el análisis se realizó únicamente sobre los roedores.

# Análisis de datos

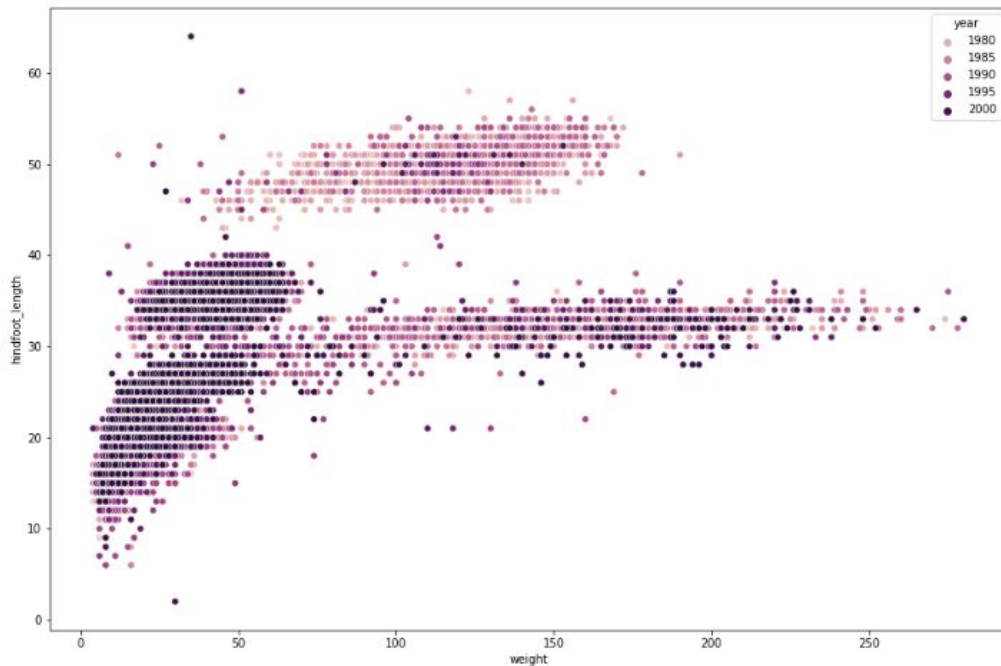
Se graficó el peso vs la longitud del contrapie de los roedores y se hizo una clasificación en colores según la especie, el sexo y el año en el cual se registró el dato.



Cuando se clasificó según la especie, se pudo distinguir los grupos que cada una de ellas forma según su peso y longitud de contrapie



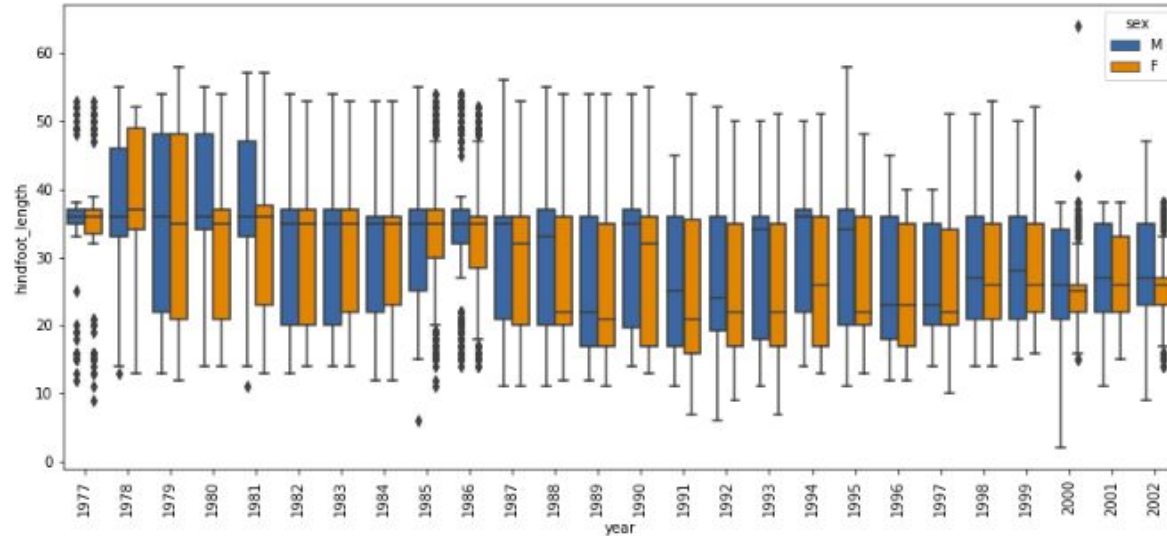
La clasificación según el sexo no muestra ninguna correlación con los datos cuantitativos.



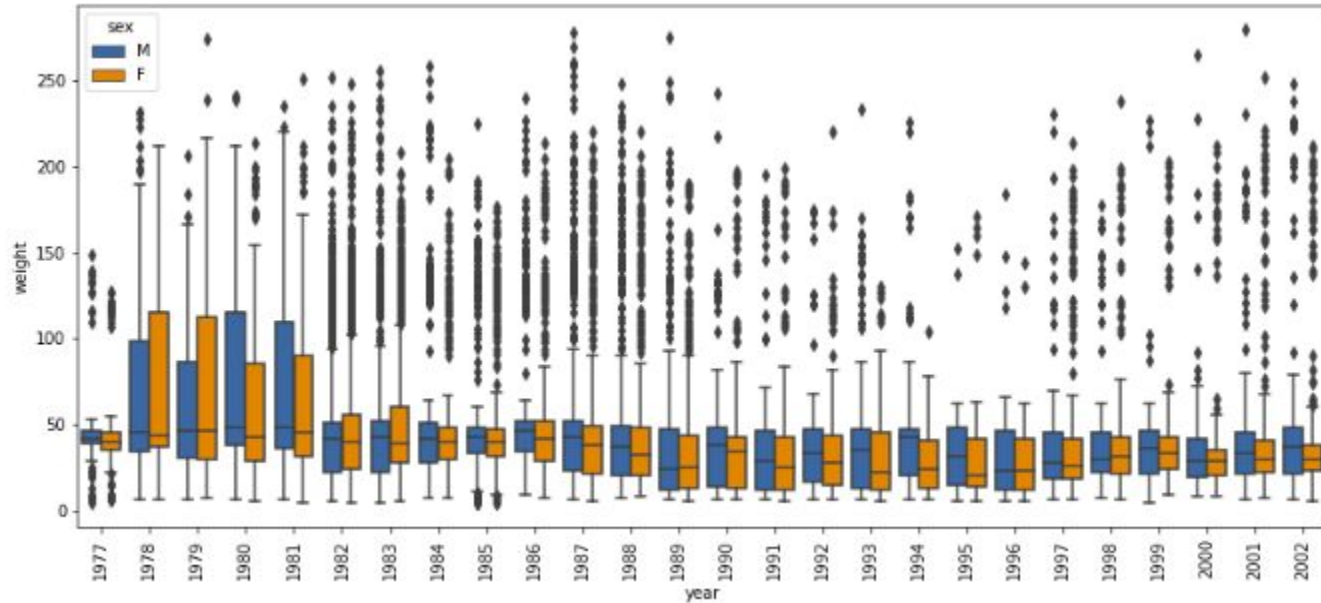
Cuando se utiliza la clasificación según el año en el que se registró el dato, nos damos cuenta que el estudio comenzó con las especies con mayor longitud de contrapíe y peso y finalizó con las especies de roedores más pequeñas.

Finalmente se realizó un diagrama de cajas y bigotes para ver si existía alguna relación entre la fisiología de las especies con su sexo, puesto que en el scatter plot no se encontró ninguna correlación.

Se realizó este diagrama para cada año ya que los gráficos anteriores nos mostraron una relación entre el año y la especie.



Analizando este gráfico, nos damos cuenta que en especies de roedores pequeñas (año 1988 en adelante), las hembras poseen en promedio una longitud de retropie menor que la de los machos. Mientras que en las especies más grandes los dos géneros tienen longitudes similares en su retropie.



Cuando realizamos el mismo diagrama para el peso, observamos un comportamiento similar al gráfico anterior. Pese a existir más datos atípicos para esta característica, nos damos cuenta que en especies cuyo peso es pequeño, las hembras tienen en promedio menor peso que los machos. Además, al igual que en el caso anterior, las especies de mayor tamaño no presentan gran diferencia en el peso entre machos y hembras.



# Conclusiones



A partir del scatterplot de peso vs longitud de retropie, se puede pensar en un algoritmo de clasificación para determinar la especie de roedor cuando solo se conocen estas dos características físicas del individuo, esto debido a la clara agrupación de los datos según su especie.

En especies de menor tamaño, las hembras tienen en promedio menor peso y longitud de retropie, siendo estas dos características para diferenciar el sexo en estas especies de roedores.

# Acceso al código



El código que contiene la transformación de la data y el análisis correspondiente se encuentra en el repositorio <https://gitmilab.redclara.net/hernandezj/ejercicios-clase-08-datos.git> en la carpeta codigo.