Herramientas computacionales: el arte de la analítica

Jorge Guerrero Díaz A01411752
José Sebastián Pedrero Jiménez A01703331
Armando Gutiérrez Rojo A01702748

Base de datos: Pokemon

La base de datos seleccionada es una que despliega información sobre 721 Pokémon, conteniendo el número de Pokédex (ID), Nombre, Número de Generación, si es legendario o no, Tipo Principal, Tipo Secundario, Total base de atributos, y el valor base de cada atributo como tal.

Lo que se espera obtener son filtros para desplegar ciertos Pokémon en función de su tipo, organizarlos por sus valores de atributo, ver si hay relaciones o tendencias por cada generación de Pokémon.

La razón por la cuál se seleccionó esta base de datos, es que a todos los integrantes del equipo les gustan los videojuegos y Pokémon. Además, se cree que la base de datos cuenta con suficientes datos numéricos y descriptivos para poder llevar las actividades de la clase a cabo.

Estadísticas Descriptivas

Con la ayuda de Jupyter, Pandas, y Python, se hizo uso de las diversas funciones que estos medios ofrecen para poder obtener valores descriptivos y con ellos, algunas conclusiones.

Datos por columna

```
#             800
Name          800
Type 1        800
Type 2        414
Total         800
HP            800
Attack        800
Defense       800
Sp. Atk       800
Sp. Def       800
Speed         800
Generation    800
Legendary     800
dtype: int64
```

Mediana de valores numéricos

```
Mediana de Total es: 450.0
Mediana de HP es: 65.0
Mediana de Attack es: 75.0
Mediana de Defense es: 70.0
Mediana de Sp. Atk es: 65.0
Mediana de Sp. Def es: 70.0
Mediana de Speed es: 65.0
```

Media de valores numéricos

```
Media de Total es: 435.1025
Media de HP es: 69.25875
Media de Attack es: 79.00125
Media de Defense es: 73.8425
Media de Sp. Atk es: 72.82
Media de Sp. Def es: 71.9025
Media de Speed es: 68.2775
```

Rango de valores numéricos

```
El mínimo valor de Total es: 180 y el máximo valor es: 780
El mínimo valor de HP es: 1 y el máximo valor es: 255
El mínimo valor de Attack es: 5 y el máximo valor es: 190
El mínimo valor de Defense es: 5 y el máximo valor es: 230
El mínimo valor de Sp. Atk es: 10 y el máximo valor es: 194
El mínimo valor de Sp. Def es: 20 y el máximo valor es: 230
El mínimo valor de Speed es: 5 y el máximo valor es: 180
```

Desviación estándar y varianzas

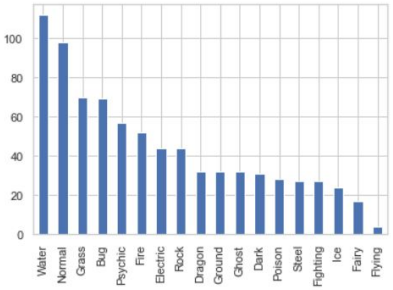
```
La desviación estándar de Total es: 119.963039755519 y su varianza es de: 14391.130907384233
La desviación estándar de HP es: 25.534669032332047 y su varianza es de: 652.0193225907373
La desviación estándar de Attack es: 32.45736586949843 y su varianza es de: 1053.4805991864816
La desviación estándar de Defense es: 31.183500559332927 y su varianza es de: 972.410707133917
La desviación estándar de Sp. Atk es: 32.72229416880157 y su varianza es de: 1070.748535669585
La desviación estándar de Sp. Def es: 27.82891579711745 y su varianza es de: 774.4485544430531
La desviación estándar de Speed es: 29.06047371716149 y su varianza es de: 844.5111326658338
```

Histogramas

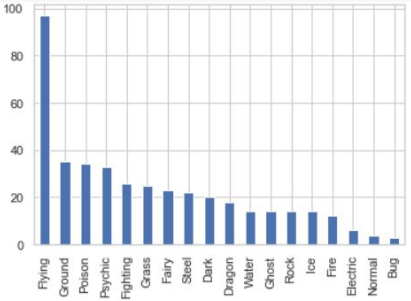
Librería matplotlib

En los histogramas podemos comparar variables que se repiten para ver cuantas incidencias hay en cierta categoría.

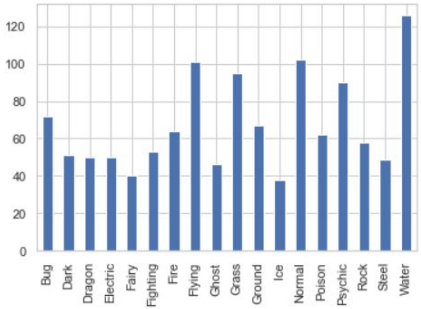
```
df["Type 1"].value_counts().plot(kind='bar')
```



Tipo Principal

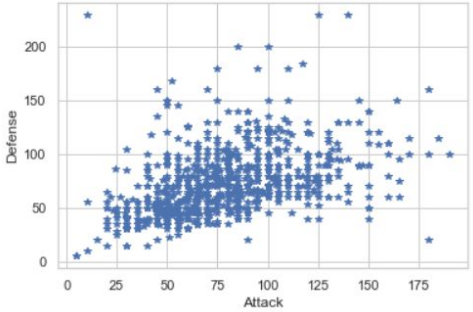


Tipo Secundario



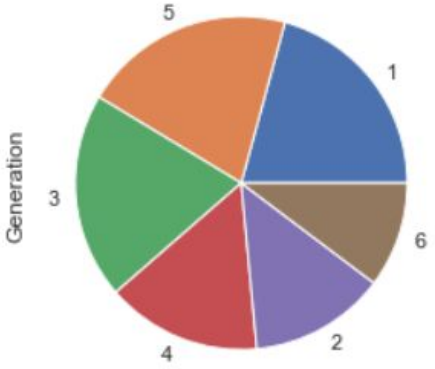
Tipo en total

```
plt.plot(df.Attack, df.Defense, '*')
plt.xlabel("Attack")
plt.ylabel("Defense")
plt.show()
```



Relación ataque-defensa

Número de Pokémon por generación




```
df['Generation'].value_counts().plot.pie();
```

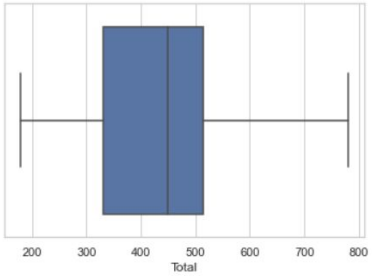
Boxplots

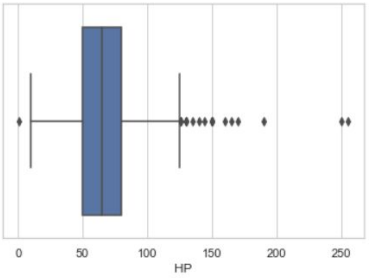
1.- Se importa la base de datos

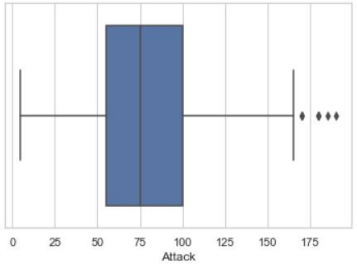
	#	Name	Type 1	Type 2	Total	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
0	1	Bulbasaur	Grass	Poison	318	45	49	49	65	65	45	1	False
1	2	Ivysaur	Grass	Poison	405	60	62	63	80	80	60	1	False
2	3	Venusaur	Grass	Poison	525	80	82	83	100	100	80	1	False
3	3	VenusaurMega Venusaur	Grass	Poison	625	80	100	123	122	120	80	1	False
4	4	Charmander	Fire	NaN	309	39	52	43	60	50	65	1	False
...	...	...	...	...	...	...	...	...	...	...	...	...	...
795	719	Diancie	Rock	Fairy	600	50	100	150	100	150	50	6	True
796	719	DiancieMega Diancie	Rock	Fairy	700	50	160	110	160	110	110	6	True
797	720	HoopaHoopa Confined	Psychic	Ghost	600	80	110	60	150	130	70	6	True
798	720	HoopaHoopa Unbound	Psychic	Dark	680	80	160	60	170	130	80	6	True
799	721	Volcanion	Fire	Water	600	80	110	120	130	90	70	6	True

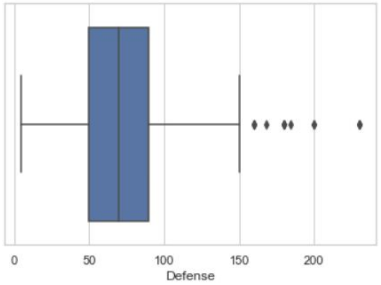
Los boxplots sirven para visualizar qué valores de cada atributo son aquellos más repetidos y los casos en donde dichos valores están por fuera de un rango.

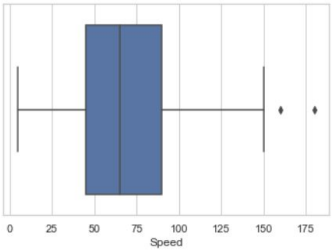
2.- Con la librería de seaborn se hacen los boxplots por cada atributo de la tabla

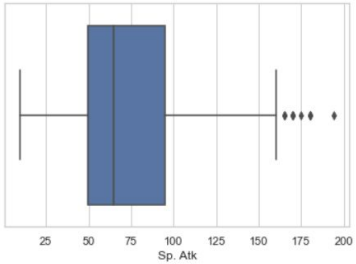


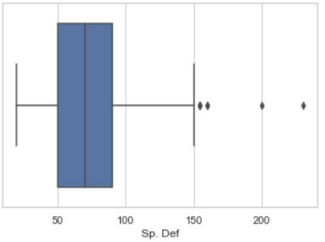




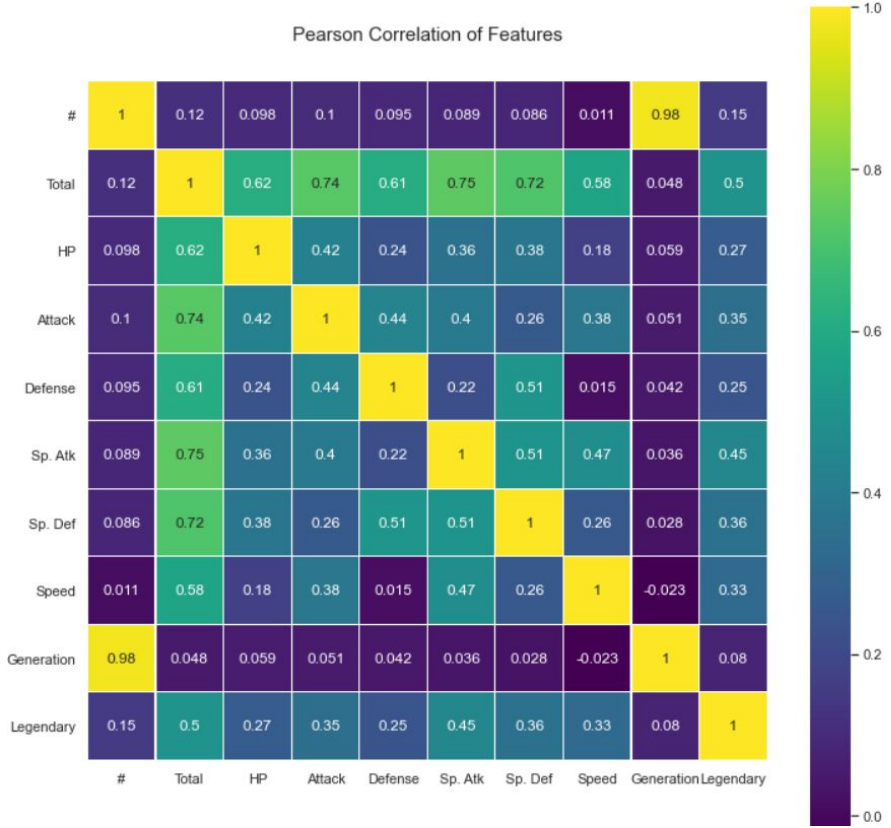








Mapas de Calor




```
colormap = plt.cm.viridis
plt.figure(figsize=(12,12))
plt.title('Pearson Correlation of Features', y=1.05, size=15)
sb.heatmap(df.corr(), linewidths=0.1, vmax=1.0, square=True,
           cmap=colormap, linecolor='white', annot=True);
```

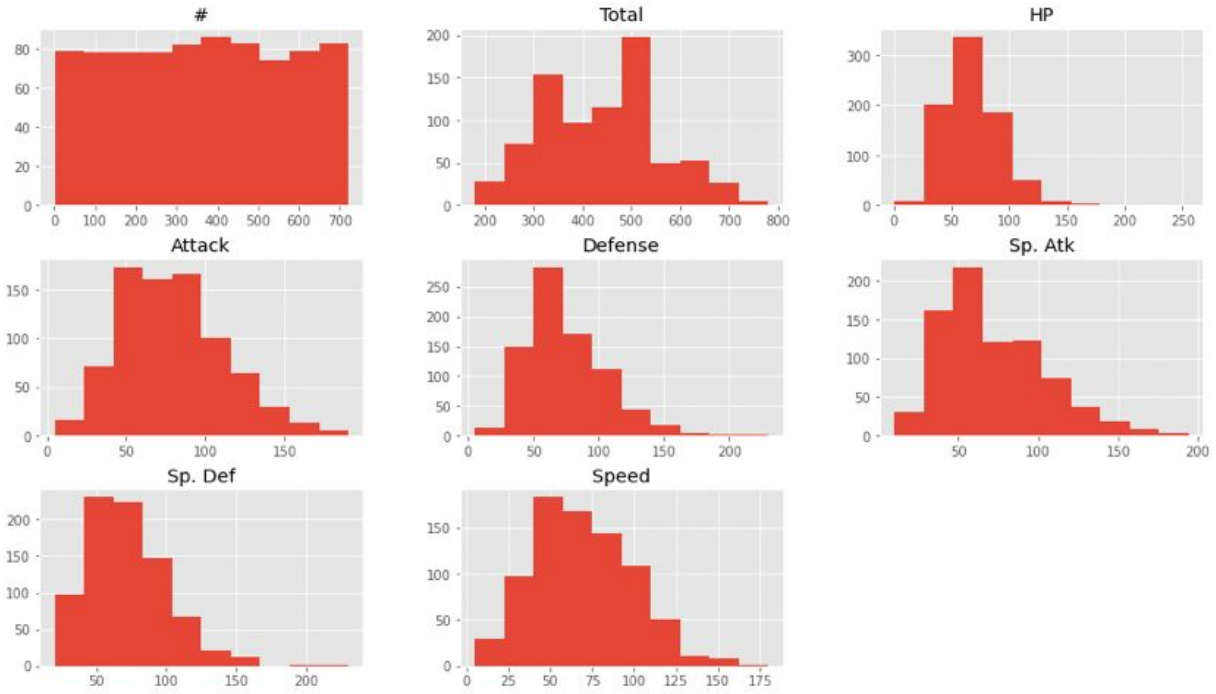
Se decidió realizar un mapa de calor para ver qué variables podrían estar relacionadas. Los resultados obtenidos indicaron que el total de atributos tiene una relación fuerte con el ataque, super ataque y súper defensa. También tiene una relación media con la vida, la defensa y la velocidad. Por otro lado, la generación tiene relación con el número del Pokémon, esto tiene sentido ya que el número incrementa conforme van surgiendo nuevos Pokémon.

K-means

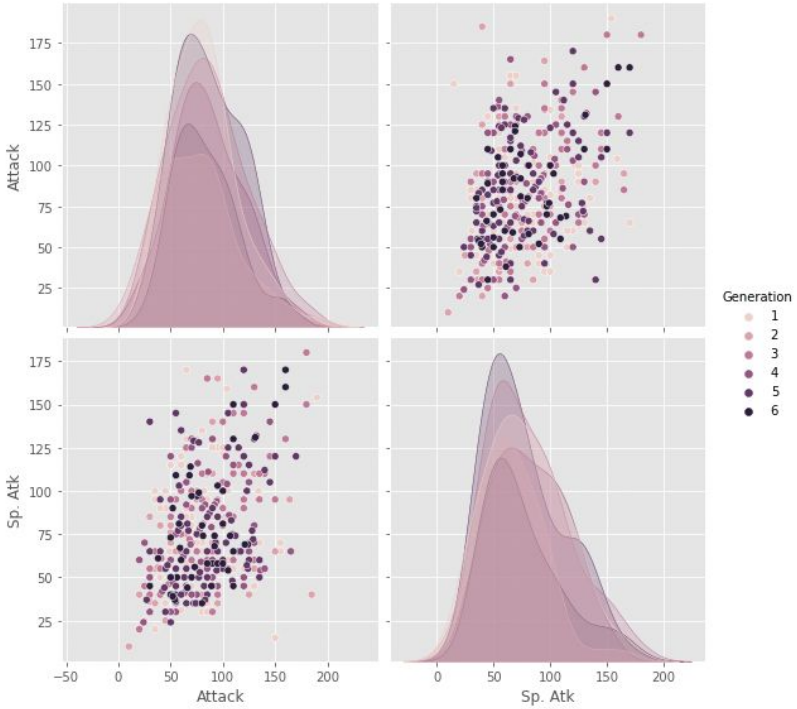
Conteo de registros

```
Generation
1    166
2    106
3    160
4    121
5    165
6     82
dtype: int64
```

Análisis gráfico



K-means

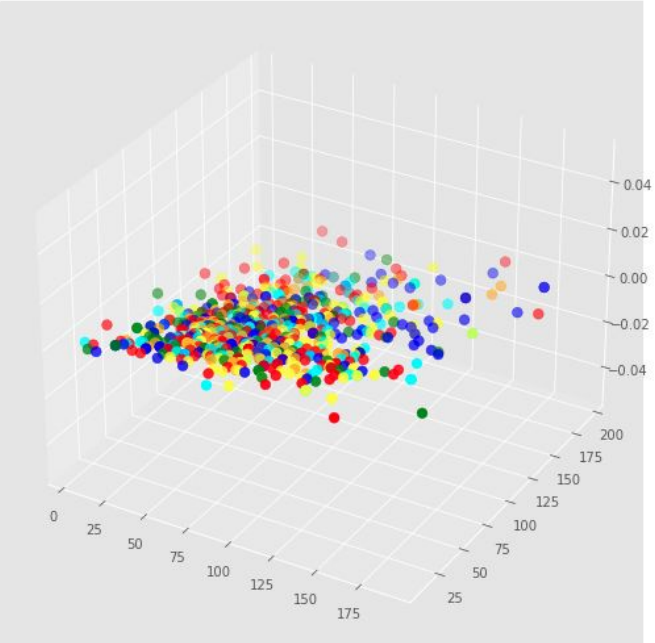


Hipótesis: Pokemon de la misma generación tienen un poder de ataque similar.

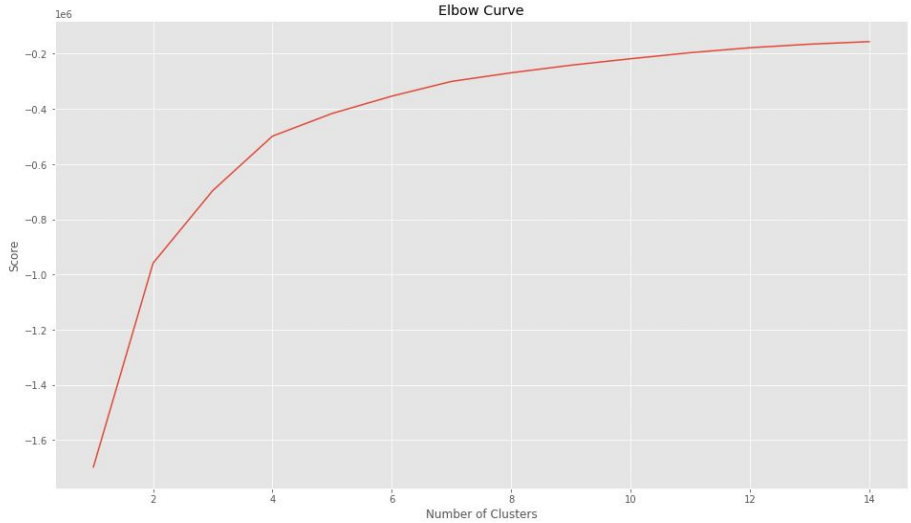
Hipótesis Nula: Pokemon de la misma generación no tienen un poder de ataque similar.

K-means

Gráfica 3D



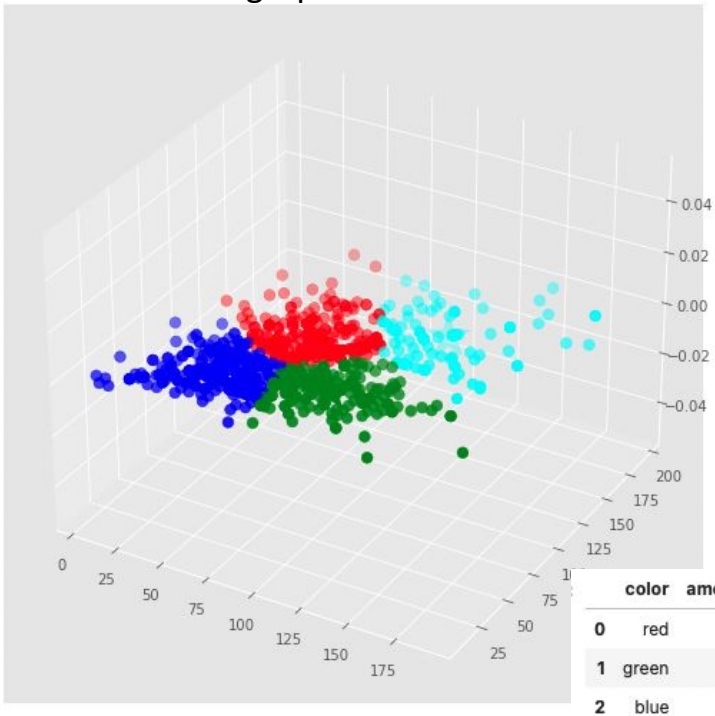
Algoritmo “punto de codo”



Tomaremos a 4 como buen valor para *K*.

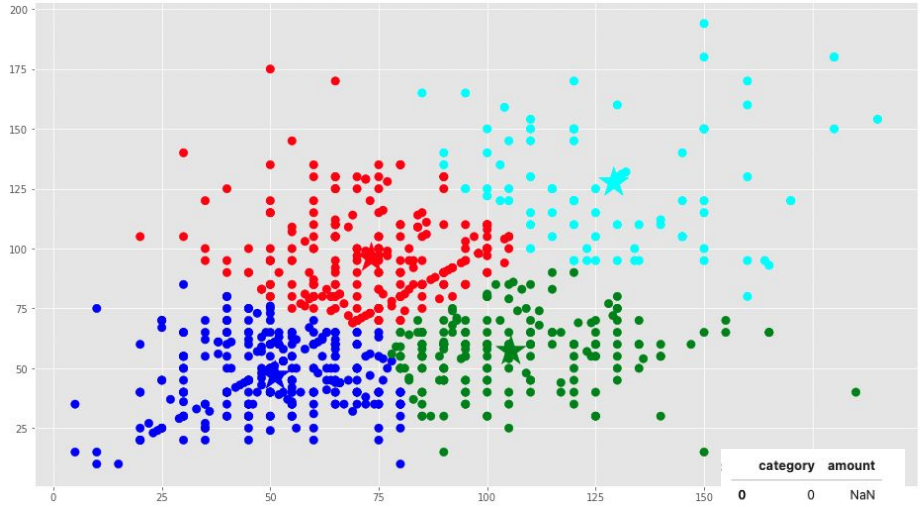
K-means

Gráfica 3D con grupos



	color	amount
0	red	243
1	green	283
2	blue	85
3	cyan	189

Gráfica 2D con grupos



Todos los centros se encuentran a una distancia muy parecida. Cada uno se encuentra a una distancia de aproximadamente 25 unidades en el eje de las X y 25 unidades en el eje de las Y del centro anterior.

	category	amount
0	0	NaN
1	1	54.0
2	2	29.0
3	3	46.0
4	4	42.0
5	5	48.0
6	6	24.0

Conclusión

Jupyter, Pandas y Python son herramientas que facilitan el análisis de datos y que nos permiten confirmar o rechazar nuestras hipótesis sobre el comportamiento del contenido de una base de datos. Además permiten realizar representaciones visuales de los datos que facilitan la interpretación de estos.