Universidad Nacional Autónoma de México

Escuela Nacional de Estudios Superiores
Unidad Morelia

Lic. en Tecnologías para la Información en Ciencias

Análisis del Dataset “Pokémon”

Estadística Descriptiva e Inferencial

Autores:

Mirell Araceli Romero Zerpa

Juan Daniel Rangel Avila

Dra. María Del Río Francos

Dr. César Andrés Torres Miranda

Mayo 2025

Índice

Resumen

En la gran franquicia de Pokémon, los distintos pokémon tienen diferentes estadísticas y características que los hacen únicos. En este proyecto abordaremos algunas de ellas, como los puntos de vida, su tipo y otras, con el objetivo de responder algunas preguntas interesantes, para conocerlos un poco mejor utilizando distintas técnicas de estadística descriptiva e inferencial. La información fue descargada de Kaggle por el usuario *jaidalmotra*. En la página se hablaba sobre los distintos tipos de datos que contenía el *dataset*.

Este conjunto de datos ha sido usado por otros estudiantes de distintas universidades con el fin de aplicar sus conocimientos en el ámbito de la estadística y para responder preguntas sobre variables cualitativas y cuantitativas.

1. Presentación de los datos

1.1. Fuente de datos

Pokémon es una franquicia que, desde su creación en 1996, ha sabido mantenerse en el mercado, aumentando cada vez más su número de criaturas y manteniéndose creativa y competitiva hasta el día de hoy. El *dataset* fue obtenido desde Kaggle, aunque también se consultó la página oficial de Pokémon, donde aparece cada uno de los Pokémon con sus distintas estadísticas, así como la explicación de cada una de las variables usadas.

También resulta interesante observar las distintas preguntas que se hace la comunidad, muchas de las cuales pueden responderse con este proyecto. Algunas de ellas son: “¿Cuál es el Pokémon más fuerte?”, “¿Cuáles son los Pokémon más parecidos de tipo fuego y agua?”, “¿Cuál es el Pokémon ideal?”, entre muchas otras que iremos abordando en el análisis.

1.2. Motivo de interés del estudio

El sistema de combate por turnos permite que cada pokémon tenga estadísticas únicas, variando así en su nombre (*name*), daño al atacar (*attack*), vida (*HP*), defensa (*defense*), velocidad (*speed*) y algunas otras. Esto hace que el sistema de combate sea dinámico y entretenido. Además, al existir tantos Pokémon, es inevitable que haya algunas estadísticas repetidas o muy parecidas, lo cual podemos analizar utilizando técnicas de estadística descriptiva para identificar ciertos patrones.

Esta gran cantidad de información ha sido diseñada por personas capacitadas en temas estadísticos y de diseño de juego, que se involucran activamente para lograr que la experiencia sea divertida. Por eso consideramos que esta es una oportunidad para entender cómo podemos responder a preguntas interesantes utilizando el dataset de Pokémon.

1.3. Presentación de las variables de un Pokémon

En el dataset original se tienen 12 columnas, las cuales se muestran a continuación.

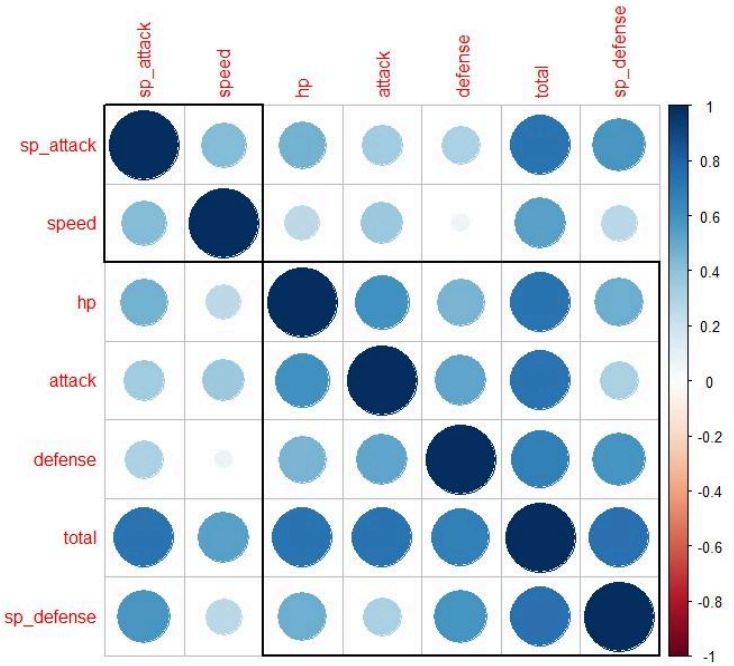
Número Consecutivo	Variable	Cantidad de datos	Tipo de variable
0	id	1072	cuantitativa discreta
1	name	1072	cualitativa nominal
2	type	1072	cualitativa nominal
3	total	1072	cuantitativa discreta
4	hp	1072	cuantitativa discreta
5	attack	1072	cuantitativa discreta
6	defense	1072	cuantitativa discreta
7	sp_sttack	1072	cuantitativa discreta
8	sp_defense	1072	cuantitativa discreta
9	speed	1072	cuantitativa discreta
10	generation	1072	cualitativa nominal
11	legendary	1072	cualitativa nominal

2. Estadística Descriptiva

Para este trabajo seleccionamos sólo las cuatro columnas que se muestran enseguida.

Número	Variable	Cantidad de datos	Tipo de variable
2	type	1072	cualitativa nominal
4	hp	1072	cuantitativa discreta
9	speed	1072	cuantitativa discreta
10	generation	1072	cualitativa nominal

Las variables escogidas se dan por un correlograma donde podemos visualizar que estos no se comportan de forma similar, es decir, tienen independencia estadística.



Ahora para no trabajar con todos los tipos de Pokémon vamos agrupar en clusters, utilizando PAM (Partitioning Around Medoids), logramos agrupar en 5 clusters.

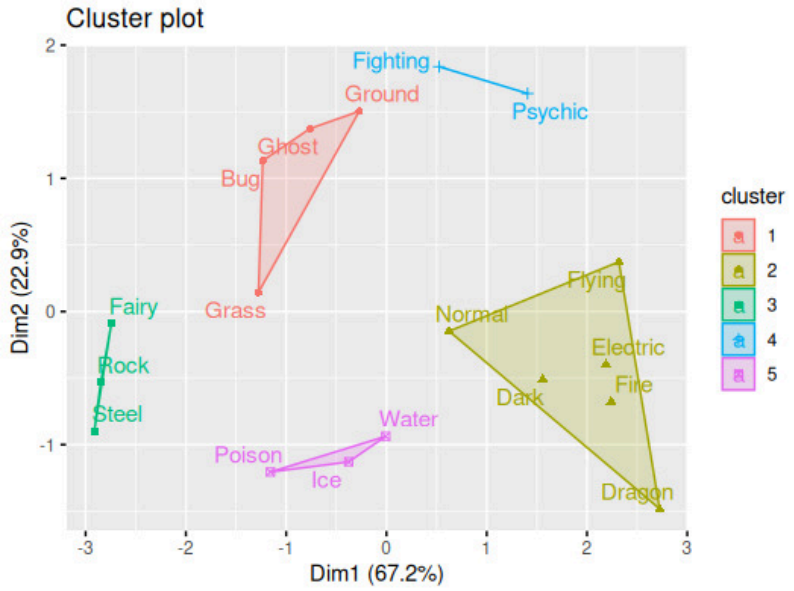
Cluster 1: Ground, Ghost, Bug, Grass

Cluster 2: Flying, Normal, Dark, Electric, Fire, Dragon,

Cluster 3: Fairy, Rock, Steel

Cluster 4: Fighting, Psychic

Cluster 5: Poison, Water, Ice



Dado que los tests de normalidad, tanto Kolmogorov-Smirnov y Shapiro-Wilk mostraron que hp y speed no siguen una distribución normal en la mayoría de clusters, excepto en el Cluster 4, como se puede ver en la siguiente tabla.

cluster	p value Shapiro	p value Kologorov
1	1.46e-6	0.0084
2	1.53e-18	0.00000343
3	3.69e-3	0.29
4	3.00e-4	0.28
5	2.24e-11	0.0358

Donde no se rechaza la normalidad, sin embargo, dado que el cluster 4 solo tiene 2 tipos no son suficientes para continuar con el análisis. Por esto **se seleccionaron los tipos Ground, Ice, Fighting, Electric y Fire** porque para estos, hp y speed son normales.

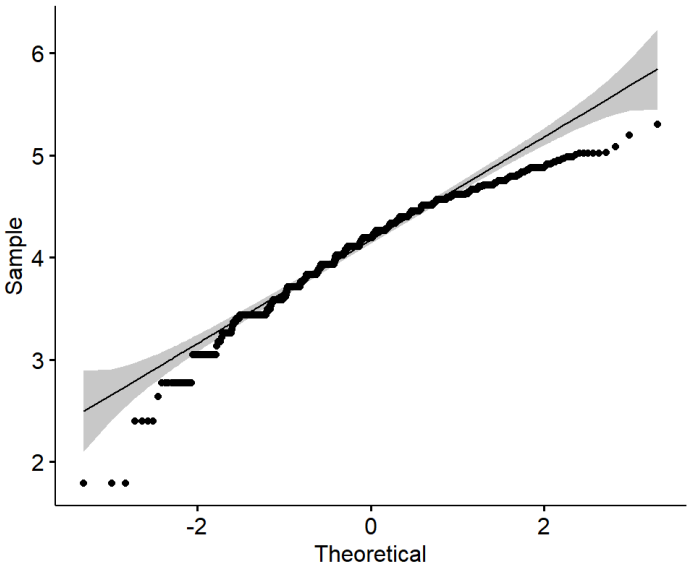
2. 1. Descripción de los valores de las variables

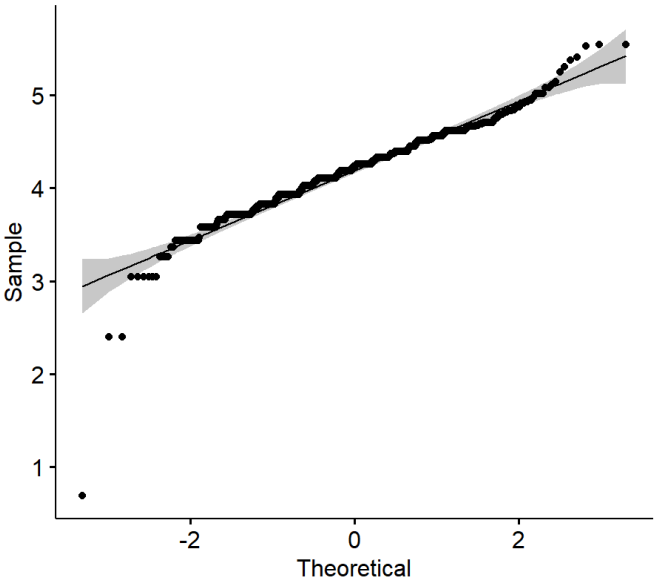
- **type.** Es útil para identificar el tipo de cada Pokémon, como agua, fuego, planta, entre otros. Este atributo nos permite responder preguntas relacionadas con los tipos, por ejemplo, si ciertos tipos tienen más vida que otros. También facilita el análisis comparativo entre grupos de tipos y sus características.
- **generation.** Indica la generación (1ª a 8ª) en la que se introdujo cada Pokémon. A lo largo de las generaciones, podemos estudiar cómo han cambiado las estadísticas base (HP, speed) entre tipos de Pokémon. Además podemos detectar sesgos de diseño, podemos hacernos preguntas como si hay tipos sobrerrepresentados en generaciones específicas.
- **speed**. Es la velocidad con la que un Pokémon ataca. Esta parte es importante de explicar, ya que tiene relación directa con la mecánica del juego. Pokémon es un juego por turnos, por lo que si ambos jugadores deciden atacar, el Pokémon con mayor velocidad atacará primero, otorgándole una ventaja sobre el oponente. En caso de que ambos tengan la misma velocidad, el orden se decide de forma aleatoria.
- **hp**. Representa la cantidad de daño que un Pokémon puede recibir antes de debilitarse, define la resistencia general del Pokémon en combate. Entre más hp, más turnos puede permanecer activo durante una batalla, lo cual puede marcar una diferencia estratégica significativa. Es decir, el valor hp representa cuánta vida tiene el Pokémon.

2.2. Resumen Estadístico de Variables

Al procesar los datos se obtuvieron los siguientes resultados.

variable	n	min	max	mean	median	sd	skewness
speed	1063	5	200	68.66	65	30.08	0.42
hp	1063	1	255	70.47	68	26.91	1.77





shapiro:

data: m1$speed_log

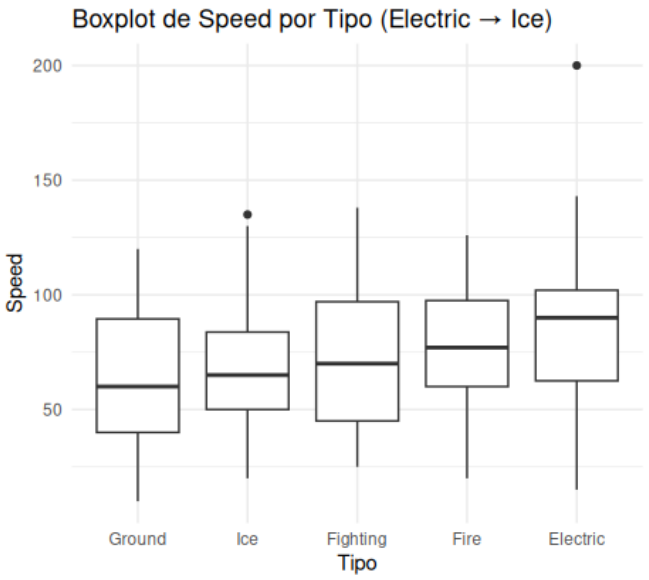
W = 0.95601, p-value < 2.2e-16

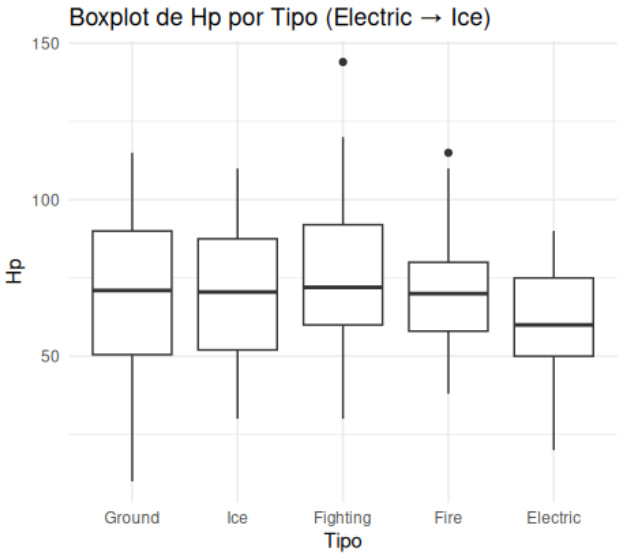
shapiro:

data: m1$hp_log

W = 0.94708, p-value < 2.2e-16

A continuación se presentan los boxplots de hp y speed de los 5 tipos de Pokémon.





2.2.1. Variables Cuantitativas

Aquí están las estadísticas descriptivas agrupados :

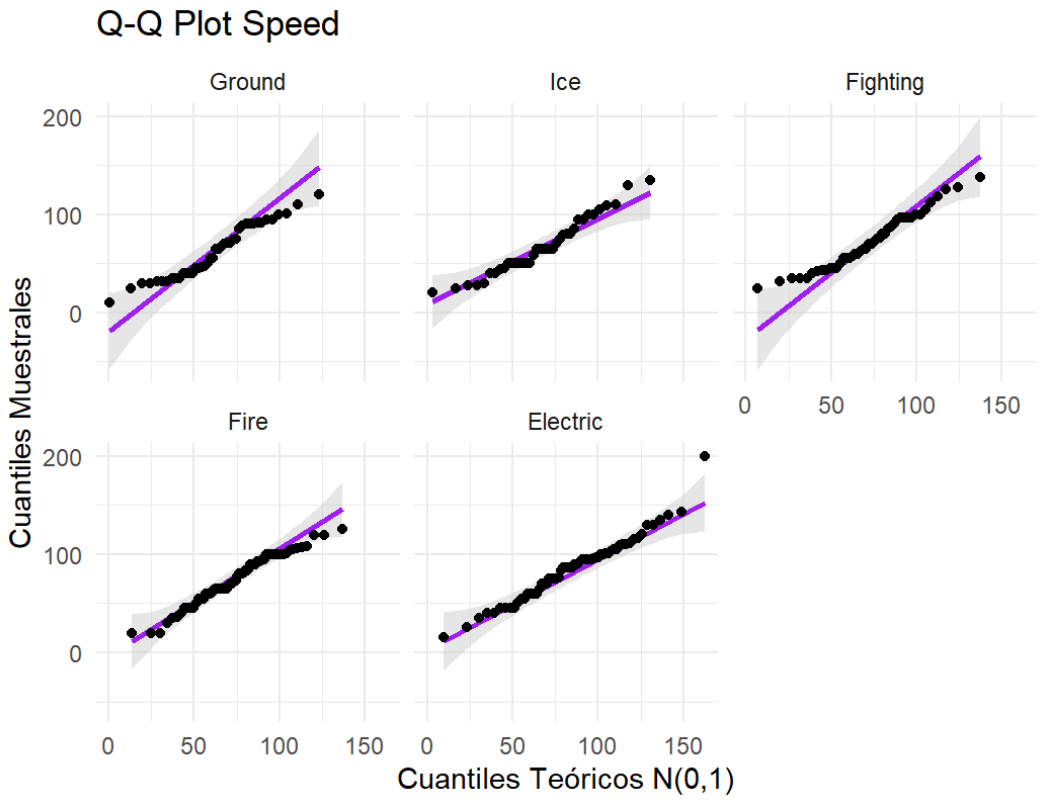
- Speed

Tipos	n	min	max	mean	median	sd	skewness
Electric	63.00	15.00	200.00	85.94	90.00	31.96	0.46
Fighting	45.00	25.00	138.00	72.11	70.00	28.72	0.40
Fire	67.00	20.00	126.00	75.34	77.00	25.45	-0.30
Ground	42.00	10.00	120.00	62.02	60.00	27.47	0.22
Ice	38.00	20.0	135.00	67.03	65.00	29.06	0.51

- Hp

Tipos	n	min	max	mean	median	sd	skewness
Electric	63.00	20.00	90.00	62.29	60.00	17.20	-0.19
Fighting	45.00	30.00	144.00	74.51	72.00	24.03	0.37
Fire	67.00	38.00	115.00	70.01	70.00	19.02	0.30
Ground	42.00	10.00	115.00	70.88	71.00	26.86	-0.27
Ice	38.00	30.00	110.0	71.68	70.50	21.67	0.05

Q-Q plots:



pruebas de normalidad:

type	Grass	W:	0.9698,	p-valor:	0.0269
type	Fire	W:	0.9707,	p-valor:	0.1151
type	Water	W:	0.9909,	p-valor:	0.5193
type	Bug	W:	0.9500,	p-valor:	0.0028
type	Normal	W:	0.9878,	p-valor:	0.3747
type	Dark	W:	0.9705,	p-valor:	0.2644
type	Poison	W:	0.9277,	p-valor:	0.0108
type	Electric	W:	0.9692,	p-valor:	0.1152
type	Ground	W:	0.9496,	p-valor:	0.0624
type	Ice	W:	0.9584,	p-valor:	0.1682
type	Fairy	W:	0.9028,	p-valor:	0.0288
type	Steel	W:	0.8880,	p-valor:	0.001
type	Fighting	W:	0.9625,	p-valor:	0.1526
type	Psychic	W:	0.9717,	p-valor:	0.073
type	Rock	W:	0.9264,	p-valor:	0.0014

type	Ghost	W:	0.9591,	p-valor:	0.1282
type	Dragon	W:	0.9659,	p-valor:	0.2521

qqplot:

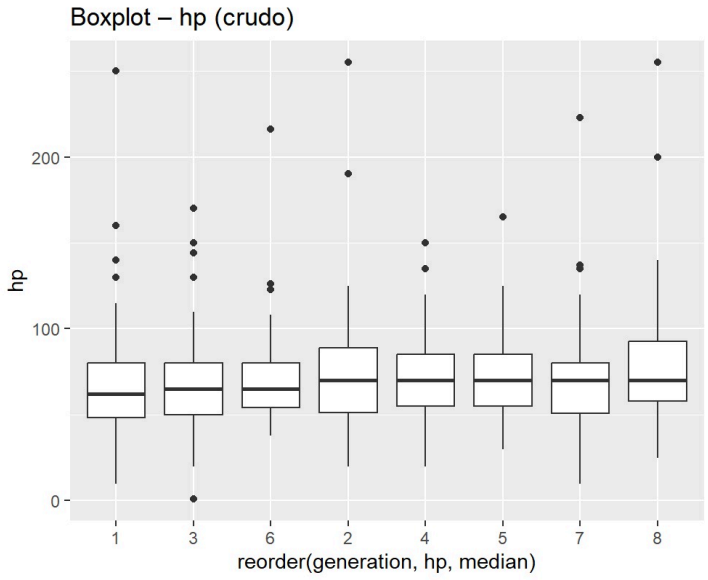
pruebas de normalidad:

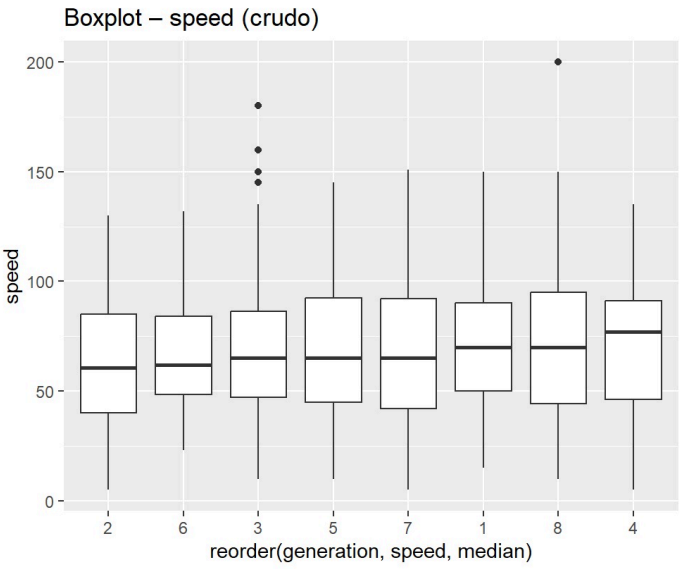
type	Grass	->	W:	0.9673,	p-valor:	0.0179
type	Fire	->	W:	0.9725,	p-valor:	0.1437
type	Water	->	W:	0.9534,	p-valor:	0.0001
type	Bug	->	W:	0.9764,	p-valor:	0.1308
type	Normal	->	W:	0.8120,	p-valor:	0
type	Dark	->	W:	0.8041,	p-valor:	0
type	Poison	->	W:	0.7094,	p-valor:	0

type	Electric	->	W:	0.9732,	p-valor:	0.1842
type	Ground	->	W:	0.9653,	p-valor:	0.2275
type	Ice	->	W:	0.9659,	p-valor:	0.2925
type	Fairy	->	W:	0.9678,	p-valor:	0.6357
type	Steel	->	W:	0.8866,	p-valor:	0.0009
type	Fighting	->	W:	0.9727,	p-valor:	0.3609
type	Psychic	->	W:	0.9386,	p-valor:	0.0008
type	Rock	->	W:	0.9762,	p-valor:	0.2907
type	Ghost	->	W:	0.7905,	p-valor:	0
type	Dragon	->	W:	0.8592,	p-valor:	0.0001

2.2.2. Variables Cualitativas

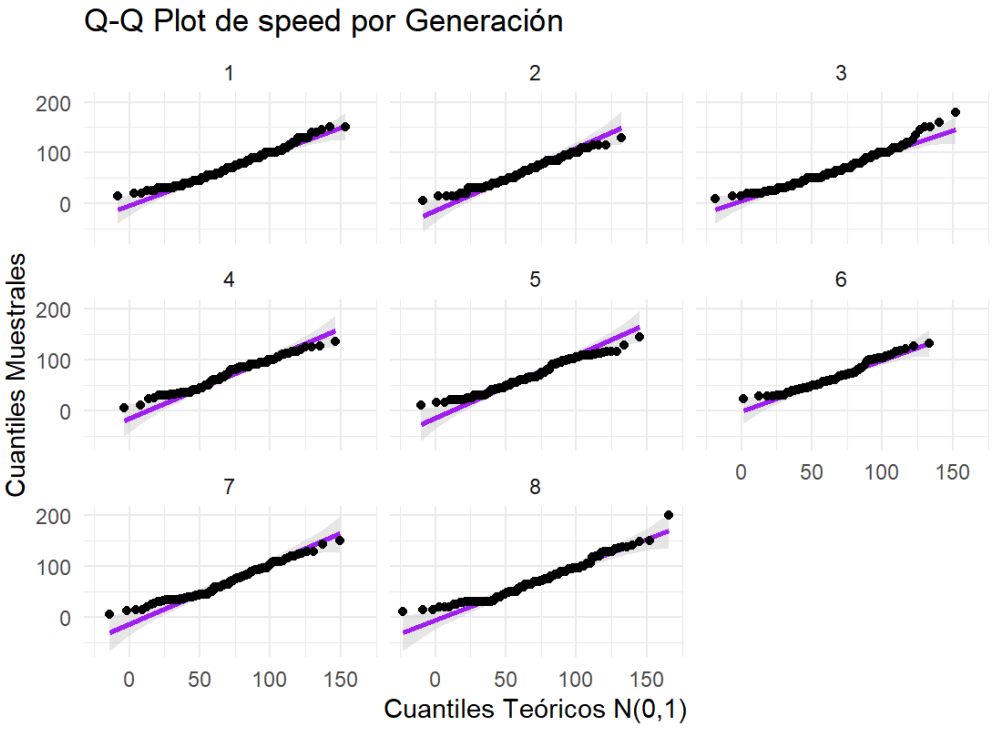
A continuación se muestra un resumen por generación de hp y speed:

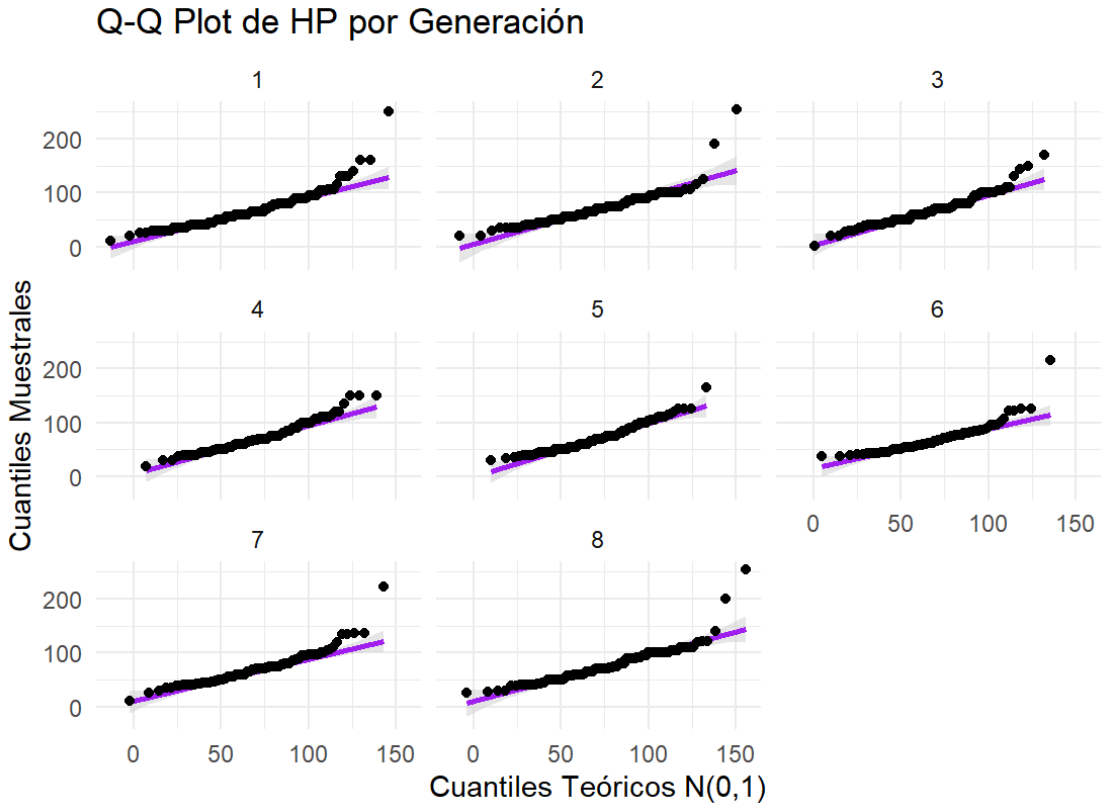




Gen	n	min	max	sd	skewness
1	178	15	150	29.18	0.37
2	106	5	130	27.26	0.18
3	160	10	180	31.33	0.75
4	121	5	135	28.48	-0.06
5	164	10	145	28.32	0.14

6	83	23	132	26.38	0.49
7	117	5	151	31.25	0.40
8	134	10	200	35.41	0.61





Pruebas de confianza para las generaciones para speed y hp:

Generación	D	P-value
1	0.0673,	0.3797
7	0.1233,	0.0521
8	0.0783,	0.3656
2	0.0827,	0.44
3	0.0984,	0.0842
4	0.1154,	0.0736
5	0.0850,	0.1766
6	0.1058,	0.2897

Generación	D	P-value
1	0.1340,	0.003
7	0.1242,	0.0496
8	0.1008,	0.1223
2	0.1205,	0.0847
3	0.1241,	0.0132
4	0.1149,	0.0755
5	0.1071,	0.0429
6	0.1266,	0.1284

3. Estadística Descriptiva

3.1. Muestreo Aleatorio

Enseguida se muestra la tabla de valores de la muestra aleatoria, para la variable generación en el atributo speed.

n	speed	n	speed	n	speed
58	**78**	**35**	**45**	**50**	**30**
145	**55**	**80**	**50**	**35**	**30**
30	**50**	**80**	**80**	**80**	**76**
71	**42**	**65**	**97**	**30**	**40**
90	**100**	**81**	**130**	**65**	**56**

3.2 Muestreo Estratificado

Para el muestreo estratificado se calculó la proporción de cada Pokémon en el dataset y al multiplicarlo por 30, nos da la cantidad de muestras que debe elegirse por cada Pokémon, para que la muestra sea representativa. Enseguida se muestra la tabla obtenida.

n	speed	n	speed	n	speed
45	**20**	**95**	**45**	**67**	**35**
50	**60**	**105**	**55**	**90**	**95**

115	45	65	40	130	65
85	48	30	90	90	95
90	56	90	50	130	50

3.3. Tabla de Frecuencias

A continuación se muestra la tabla de la muestra aleatoria.

	Speed	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada	Z
0	30	4	0.1333	0.1333	-1.238557
1	35	2	0.0667	0.2	-1.062292
2	40	1	0.0333	0.2333	-0.886027
3	42	1	0.0333	0.2667	-0.815521
4	45	1	0.0333	0.3	-0.709762
5	50	3	0.1	0.4	-0.533496
6	55	1	0.0333	0.4333	-0.357231
7	56	1	0.0333	0.4667	-0.321978
8	58	1	0.0333	0.5	-0.251472
9	65	2	0.0667	0.5667	-0.0047
10	71	1	0.0333	0.6	0.206818
11	76	1	0.0333	0.6333	0.383083
12	78	1	0.0333	0.6667	0.453589
13	80	4	0.1333	0.8	0.524095
14	81	1	0.0333	0.8333	0.559349
15	90	1	0.0333	0.8667	0.876626
16	97	1	0.0333	0.9	1.123397
17	100	1	0.0333	0.9333	1.229157
18	130	1	0.0333	0.9667	2.286748
19	145	1	0.0333	1	2.815544

A continuación se muestra la tabla de la muestra estratificada.

	Speed	Frecuencia absoluta	Frecuencia relativa	Frecuencia relativa acumulada	Z
0	20	1	0.0333	0.0333	-1.76591
1	30	1	0.0333	0.0667	-1.418746
2	35	1	0.0333	0.1	-1.245164
3	40	1	0.0333	0.1333	-1.071581
4	45	3	0.1	0.2333	-0.897999
5	48	1	0.0333	0.2667	-0.79385
6	50	3	0.1	0.3667	-0.724417
7	55	1	0.0333	0.4	-0.550834
8	56	1	0.0333	0.4333	-0.516118
9	60	1	0.0333	0.4667	-0.377252
10	65	2	0.0667	0.5333	-0.20367
11	67	1	0.0333	0.5667	-0.134237
12	85	1	0.0333	0.6	0.490659
13	90	5	0.1667	0.7667	0.664242
14	95	3	0.1	0.8667	0.837824
15	105	1	0.0333	0.9	1.184988
16	115	1	0.0333	0.9333	1.532153
17	130	2	0.0667	1	2.0529
18	130	1	0.0333	0.9667	2.286748
19	145	1	0.0333	1	2.815544

3.4. Tablas de tendencia central, asimetría y dispersión.

Para la muestra aleatoria:

	Media	Mediana	Moda	Desviación estándar	Varianza	Mínimo	Máximo	Asimetría
speed	65.13	61.5	30	28.85	832.4	30	145	0.9

Para la muestra estratificada:

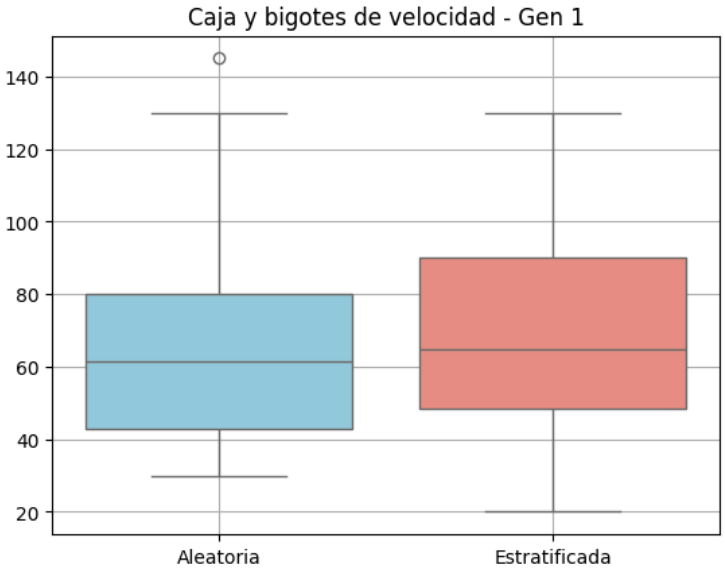
	Media	Mediana	Moda	Desviación estándar	Varianza	Mínimo	Máximo	Asimetría
speed	70.87	65	90	29.3	858.33	20	130	0.34

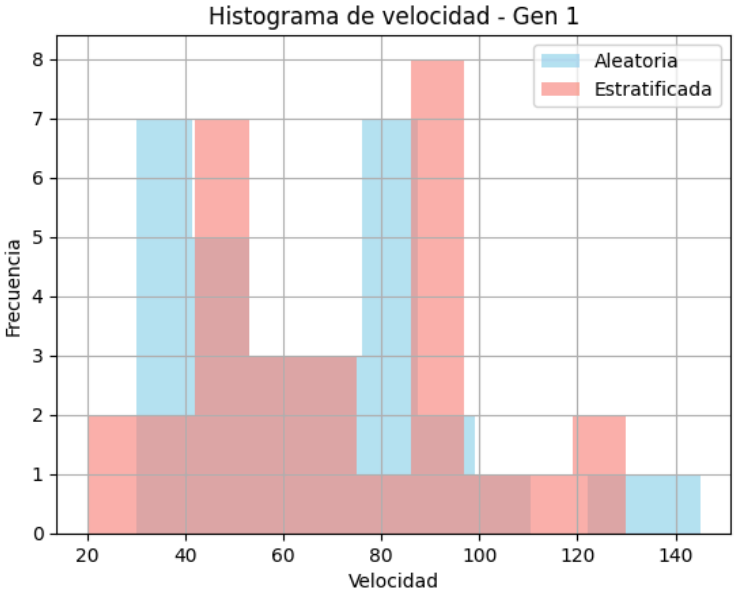
Análisis: la media de la muestra estratificada (70.87) tiene un valor más cercano a la media de la población (72.71). La mediana de la muestra estratificada (65) es más cercana a la mediana de la población (70). La moda de la muestra estratificada (90) es igual a la moda poblacional. En el caso de la desviación estándar vemos que la muestra estratificada (29.3) está más cerca al valor de la población (29.18).

3.5. Análisis.

Se puede inferir que la muestra estratificada es más representativa de la población, comparado con la muestra aleatoria.

3.6. Diagramas e Histogramas





3.7. Probabilidad de la media

Nivel de confianza: 85%	**Nivel de confianza: 95%**	**Nivel de confianza: 99%**
Hipótesis aleatoria:	**Hipótesis aleatoria:**	**Hipótesis aleatoria:**

Media real Gen 1: 72.71	**Media real Gen 1: 72.71**	**Media real Gen 1: 72.71**
Media aleatoria: 65.13	**Media aleatoria: 65.13**	**Media aleatoria: 65.13**
Valor p: 0.1611	**Valor p: 0.1611**	**Valor p: 0.1611**
Conclusión: No se rechaza H0	**Conclusión: No se rechaza H0**	**Conclusión: No se rechaza H0**
Hipótesis estratificada:	**Hipótesis estratificada:**	**Hipótesis estratificada:**
Media real Gen 1: 72.71	**Media real Gen 1: 72.71**	**Media real Gen 1: 72.71**
Media estratificada: 70.87	**Media estratificada: 70.87**	**Media estratificada: 70.87**
Valor p: 0.7332	**Valor p: 0.7332**	**Valor p: 0.7332**
Conclusión: No se rechaza H0	**Conclusión: No se rechaza H0**	**Conclusión: No se rechaza H0**

3.8 ¿Se parece la media muestral a la media poblacional?

En este punto hicimos una prueba de hipótesis para comparar la media de velocidad de cada muestra (la aleatoria y la estratificada) con la media real de todos los Pokémon de la Generación 1. En ambos casos, el valor p fue mayor a 0.15, así que no se rechaza la hipótesis nula. Esto significa que la media de cada muestra es la misma que la media poblacional.

En pocas palabras, las dos muestras representaron bien la media real. También observamos que el valor p es mayor en la muestra estratificada, lo que indica que la probabilidad de la hipótesis nula es mayor. Es decir, la media muestral y la media poblacional son iguales (con el nivel de significancia dado).

4. Análisis Bivariado

4.1. Pares significativos:

Se encontraron 5 pares significativos:

pares significativos	var1	var2	p_val
1	hp	speed	6.10E-17

2	defense	hp	**3.49E-55**
3	attack	speed	1.88E-32
4	attack	hp	**3.55E-101**
5	attack	defense	**1.15E-74**

4.2. Centroide, covarianza muestral y de Pearson

a) **Centroide:** Muestra que los Pokémon tienen en promedio

speed	hp	defense	attack
68.65757	70.46754	75.01976	80.97648

b) **Covarianza muestral:** Muestra cómo varían conjuntamente las variables

	speed	hp	defense	attack
speed	904.592614	141.3119	2.479464	356.023
hp	141.311851	724.3735	251.745934	397.43
defense	2.479464	251.7459	976.665334	466.9995
attack	356.023013	397.43	466.999523	1057.4712

Podemos interpretar lo siguiente:

- En la diagonal principal (varianzas):
    - Attack tiene la mayor variabilidad (1057.47), seguido de defense (976.67).
- Covarianzas fuera de la diagonal:
    - Speed-Defense (2.48): Casi independientes (covarianza cercana a 0).
    - HP-Defense (251.75): Relación positiva.
    - Attack-Defense (467.00): Fuerte covariación positiva, lo que nos dice que P**okémon con mucho ataque tienden a tener defensas altas**.
    - Speed-Attack (356.02): Covariación positiva notable.

c) **Coeficiente de pearson (para speed y hp)**

0.1745705 es una correlación positiva débil

De acuerdo a esto, los Pokémon más rápidos tienden ligeramente a tener más HP, pero la relación es casi casi nula .

4.3. Análisis de Regresión Lineal: Speed y HP

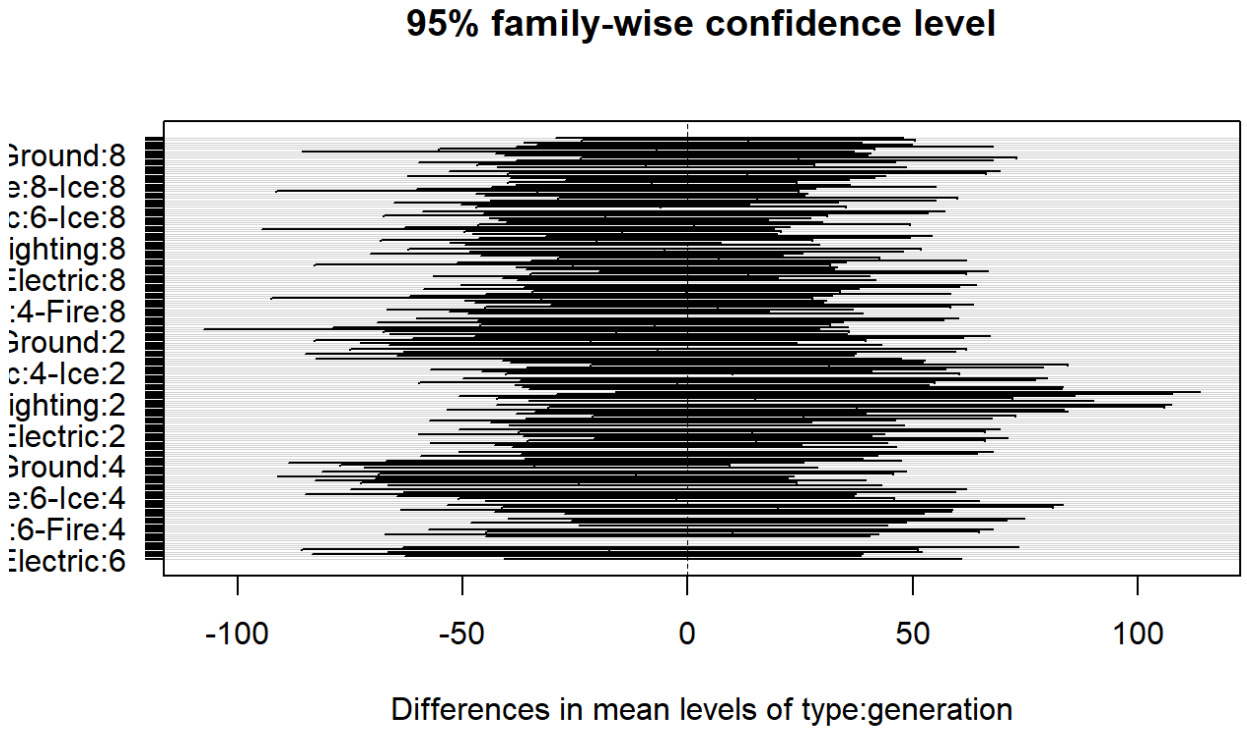
Al ejecutar el código, nos dio la siguiente salida:

```
lm(formula = speed ~ hp, data = m1)
Residuals:
    Min      1Q  Median      3Q     Max
-76.247 -23.261  -0.786  21.335 129.483

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 54.91066    2.54800  21.551  < 2e-16 ***
hp           0.19508    0.03378   5.775 1.01e-08 ***

Signif. codes:
0 ‘’ 0.001 ‘’ 0.01 ‘’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.63 on 1061 degrees of freedom
Multiple R-squared:  0.03047,   Adjusted R-squared:  0.02956
F-statistic: 33.35 on 1 and 1061 DF,  p-value: 1.01e-08
```

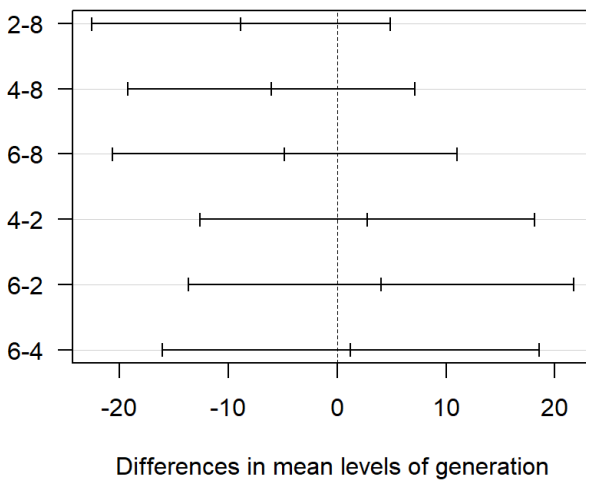
El análisis de regresión lineal entre HP y Speed mostró una relación positiva estadísticamente significativa ($\beta = 0.195$, $p < 0.001$), pero explica muy poco ($R^2 = 3.05\%$) ya que solo el 3.05% de la variabilidad en Speed es explicada por HP. La ecuación de regresión (Speed=54.91+0.195(HP)) estima que, por cada punto adicional de HP, la velocidad aumenta 0.195 en promedio, con un error residual de ±29.63 unidades. Bajo R^2 HP no es un predictor suficiente para la velocidad. Por lo que otras variables influyen más en la velocidad que solo hp.

5. Anovas:

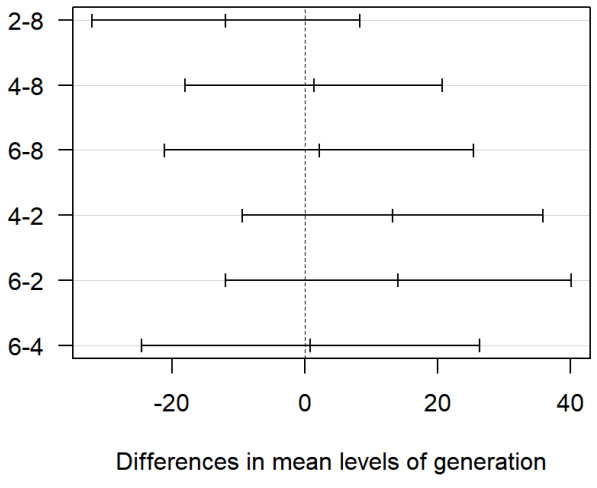
type con generación:



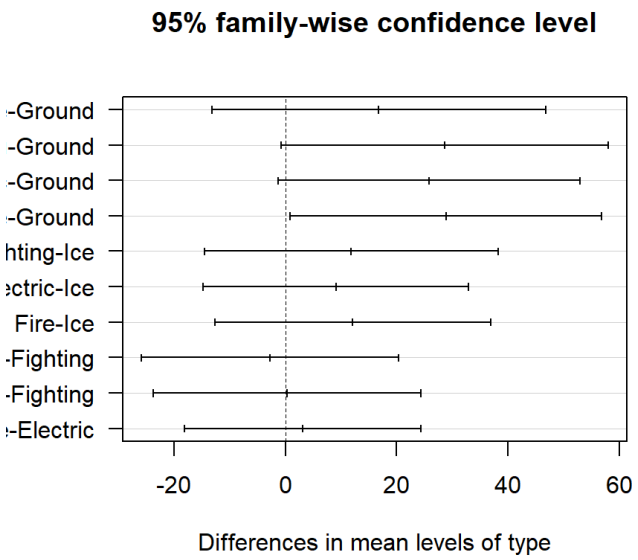
hp generación:



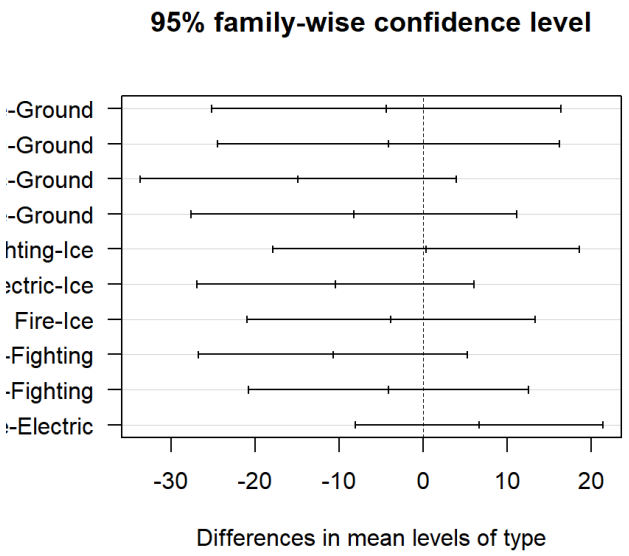
speed generación:



speed type:



hp type:



6. Referencias

https://www.kaggle.com/datasets/jaidalmotra/pokemon-dataset?select=Pokemon.csv

Pokédex | Pokemon.es

Projeto Pokemon Ideal