CONOCIENDO LOS DATOS

Oscar Daniel Rodriguez Gallego

1. **Conocer las fuentes de datos:**
   * ¿Cuántas fuentes de datos existen, ¿cuál es su tipo y su origen?
   * ¿Qué información contiene cada fuente de datos y cómo se relacionan entre sí?

Biometría. Se refieren a la identificación automática de una persona en base a sus características anatómicas o personales, como por ejemplo una firma biométrica. Hablamos tanto de reconocimiento facial como de genética (ADN).

Máquina a máquina. Se refiere al Internet de las Cosas, son estas tecnologías las que permiten la conexión de diferentes dispositivos entre sí. Un ejemplo es el GPS, pero también el llamado chip NFC (esta tecnología se basa en la comunicación inalámbrica y permite transmitir datos de forma segura: básicamente integrados en smartphones y tablets). Todo un mundo de posibilidades también se puede encontrar en parquímetros, cajeros automáticos, máquinas expendedoras...

Datos transaccionales. Los datos registrados en los servicios de pago son parte de las operaciones normales que ocurren en las transacciones normales. También existen call centers, mensajería, solicitud, presentación y registro de documentos y los generados a partir de pagos con tarjeta, pagos en línea.

hecho por el hombre. Todos estos registros para operadores de atención al cliente: call center, correo electrónico o historia clínica electrónica.

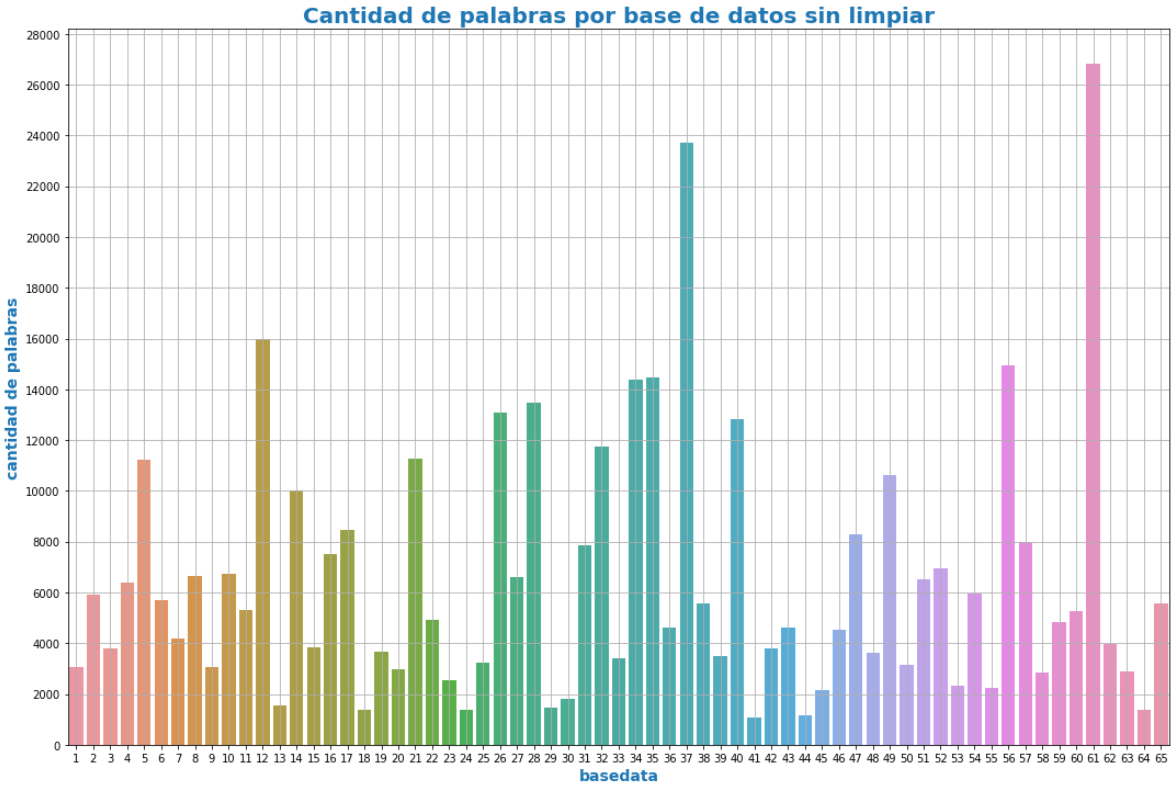
Web y redes sociales. Son cosas que vienen de la red y según los expertos, la mayor parte del pastel se llama Big Data y es una de las fuentes de datos más utilizadas en la actualidad. Estamos hablando de información generada al hacer clic en enlaces y elementos. Pero todo está incluido en las búsquedas que hacemos, por ejemplo, en Google, publicaciones en redes sociales (Twitter, Facebook, Linkedin…) y contenido web como páginas, enlaces o imagen de imágenes.

* ¿Cómo conocer la calidad de la información que contiene cada fuente?

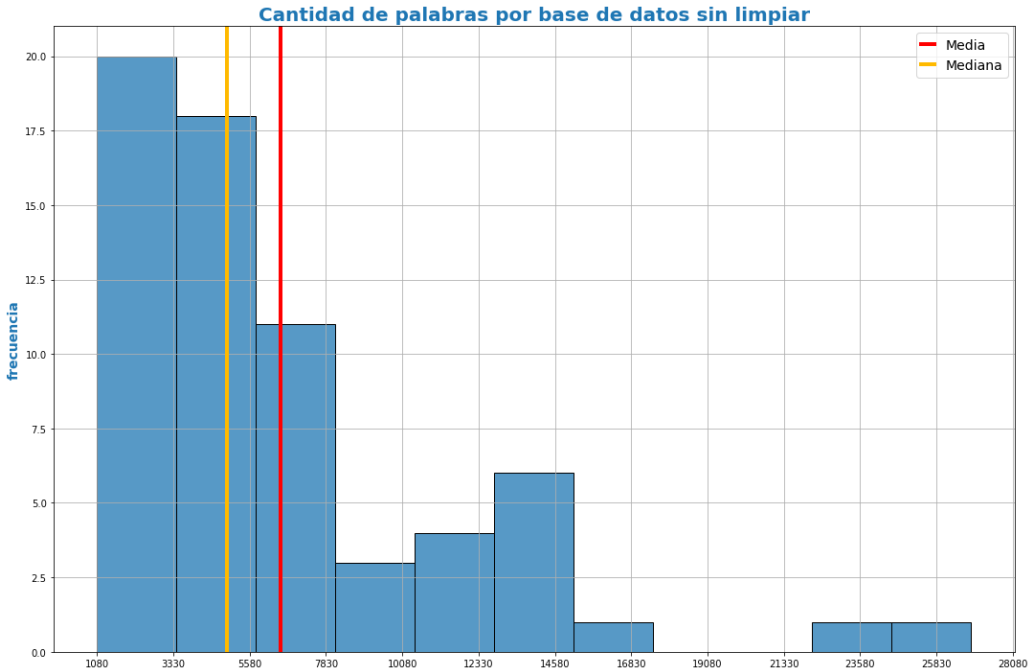
Calidad de datos es la cualidad de un conjunto de información recogida en una base de datos, un sistema de información o un data warehouse que reúne entre sus atributos la exactitud, completitud, integridad, actualización, coherencia, relevancia, accesibilidad y confiabilidad necesarias para resultar útiles al procesamiento, análisis y cualquier otro fin que un usuario quiera darles.

2. **Muestreo de datos:**
   Realizar un muestreo de cada una de las diferentes fuentes de información obtenidas en la recolección de la búsqueda de las palabras.

Para empezar, se recibió varias fuentes de datos en archivos planos. En total son 65 archivos txt. A continuación, se muestra una gráfica de la cantidad de palabras que tiene cada base de datos, cada valor corresponde al orden de las bases de datos entregadas.



Después de eso se observó la distribución de las palabras y se determino que tiene un sesgo a la derecha y además de las 65 bases de datas.

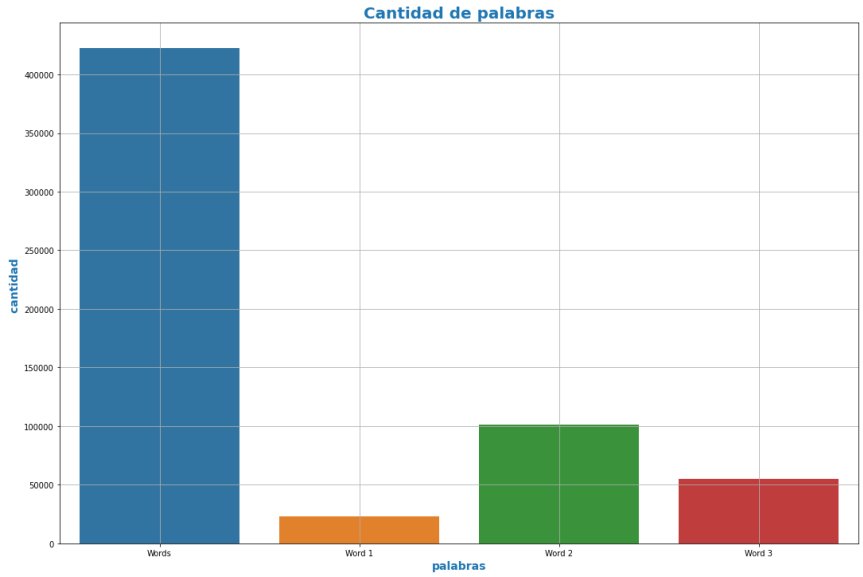


De la gráfica anterior se obtuvo que en promedio hay 6505.03 palabras y la mediana es de 4912 palabras en las bases de datos.

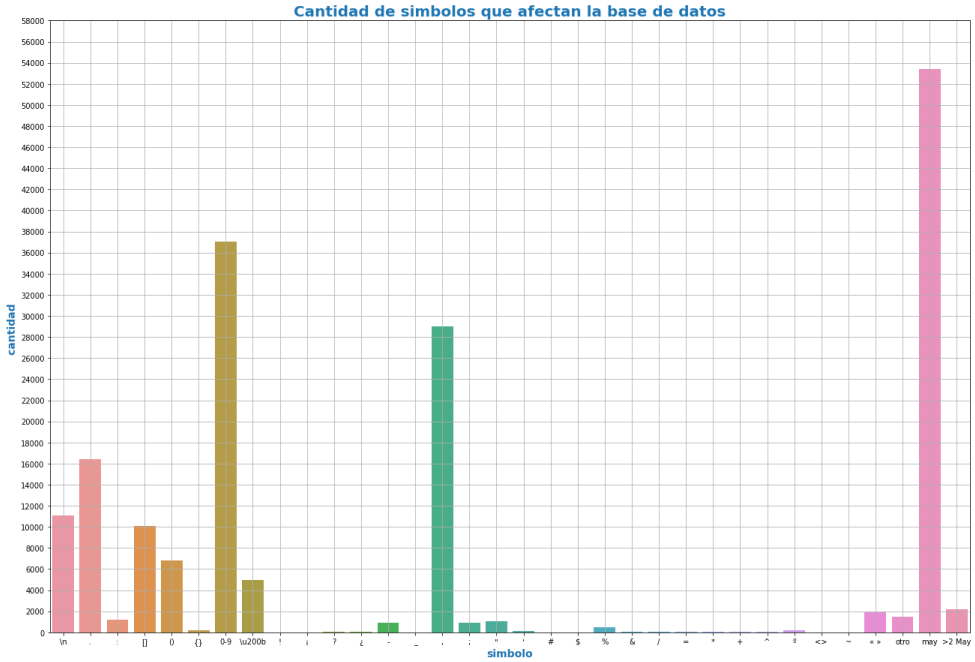
3. Recolección inicial de datos:

* Realizar la identificación y unificación de las fuentes de información obtenidas en la recolección de la búsqueda de las palabras.
* Genere un reporte con la información obtenida de la recolección de los datos

Después de unir las bases de datos se obtuvo que hay 422827 palabras de las cuales alrededor de 178574 son palabras menores de 3 caracteres.



Por otro lado, se identifico que la base de datos tiene bastante símbolos que se deben limpiar, a continuación, se muestra una grafica con la cantidad de símbolos aproximada que se eliminaron de la base de datos.

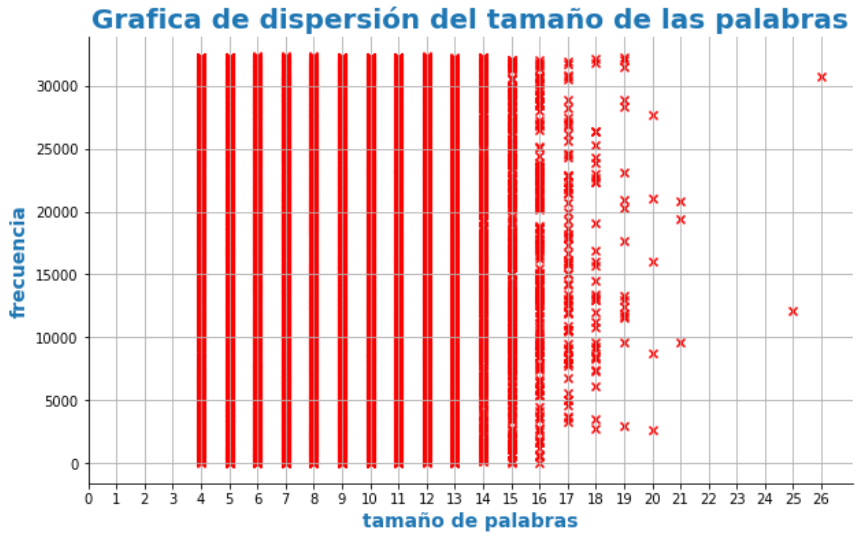


Por otra parte, la base de datos contiene 685 caracteres, aproximadamente 60 corresponde al alfabeto español y los demás se dividen en operadores aritméticos, letras griegas, latinas, números etc. Cuando se eliminó las stop words, palabras repetidas y palabras menores a 3 caracteres y eliminar las palabras con símbolos diferentes al español se llegó a 32323 palabras.

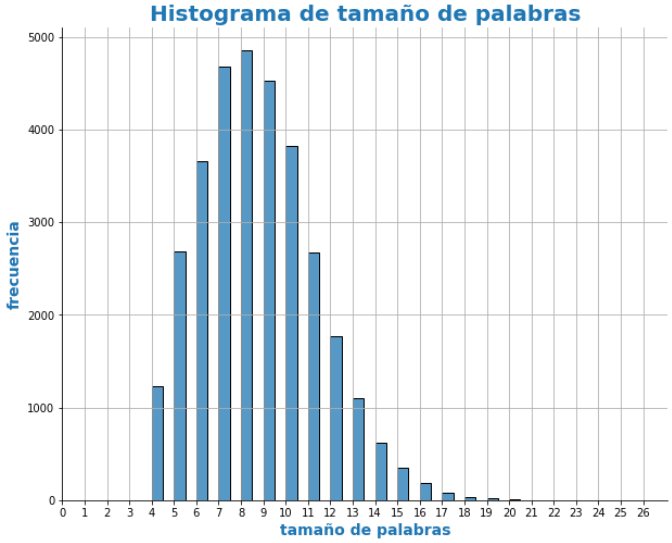
4. Medidas de tendencia central:

* Determinar la media, moda y mediana de las longitudes de caracteres de las palabras disponibles. Adicionalmente puede realizar análisis de tendencia central en otras variables si lo desea.

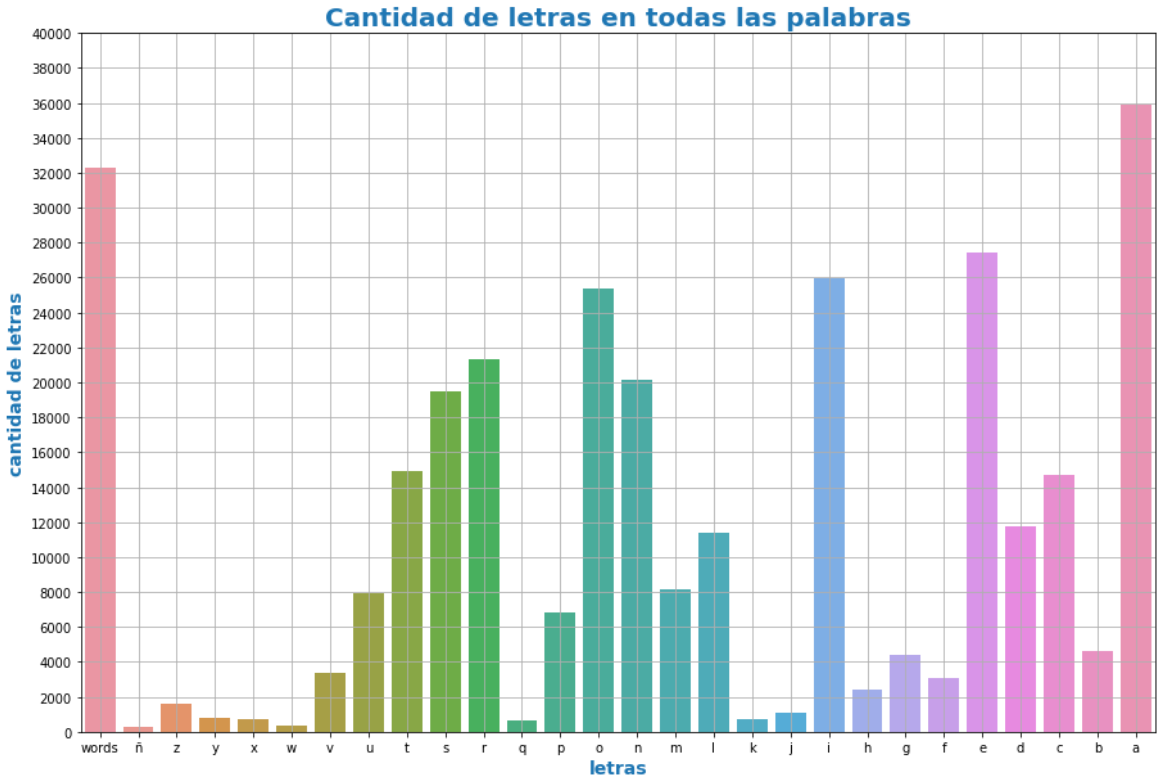
Se hizo un gráfico de dispersión de las palabras, sin embargo, proporciona poca información.



Para obtener un grafico de mejor calidad, se hizo un histograma para conocer cómo se comportaban los datos limpiados.



Como se puede los datos tienden a tener un comportamiento normal, sin embargo, como no ha palabras con menos de 4 caracteres, la media esta un poco por arriba de la mediana. Con lo anterior se calculó la media medina y moda: 8.52, 8, 8 respectivamente. Por ultimo se miro la cantidad de repeticiones que cada letra tiene en las diferentes palabras, y se observó que las vocales son las letras que más se repiten.



5. Medidas de posicionamiento estadístico:

* Determinar los cuartiles 1, 2 y 3 y percentil 90 para la distribución de la longitud de las palabras presentadas en el juego.
* Determinar los cuartiles 1, 2 y 3 para la distribución del posicionamiento de cada una de las letras del alfabeto en las palabras presentadas en el juego.

A continuación, se muestra una tabla con los diferentes cálculos:

Descripción	Percentil 90	Cuartil 3	Cuartil 2	Cuartil 1
ñ	6.0	5.0	4.0	2.0
z	9.0	7.0	5.0	3.0
y	7.0	5.0	4.0	2.0
x	4.0	2.0	1.0	1.0
w	5.0	4.0	2.0	0.0
v	7.0	5.0	3.0	1.0
u	6.0	4.0	3.0	1.0
t	8.0	6.0	4.0	2.0
s	10.0	8.0	5.0	2.0

r	7.0	5.0	4.0	2.0
q	5.0	4.0	2.0	1.0
p	5.0	3.0	2.0	0.0
o	9.0	7.0	5.0	2.0
n	9.0	7.0	5.0	2.0
m	7.0	5.0	2.0	1.0
l	8.0	5.0	3.0	2.0
k	7.0	4.0	3.0	0.0
j	5.0	4.0	2.0	0.0
i	8.0	6.0	4.0	3.0
h	6.0	4.0	2.0	0.0
g	6.0	5.0	3.0	2.0
f	6.0	4.0	2.0	0.0
e	8.0	6.0	3.0	1.0
d	9.0	7.0	5.0	2.0
c	8.0	6.0	3.0	0.0
b	7.0	5.0	2.0	1.0
a	9.0	7.0	5.0	2.0

6. Medidas de dispersión estadística:

* Determinar el rango, varianza y desviación estándar para la longitud de las palabras presentadas en el banco de palabras.
*Determinar el rango, varianza y desviación estándar para la posición de cada letra dentro de una palabra en el banco de palabras.

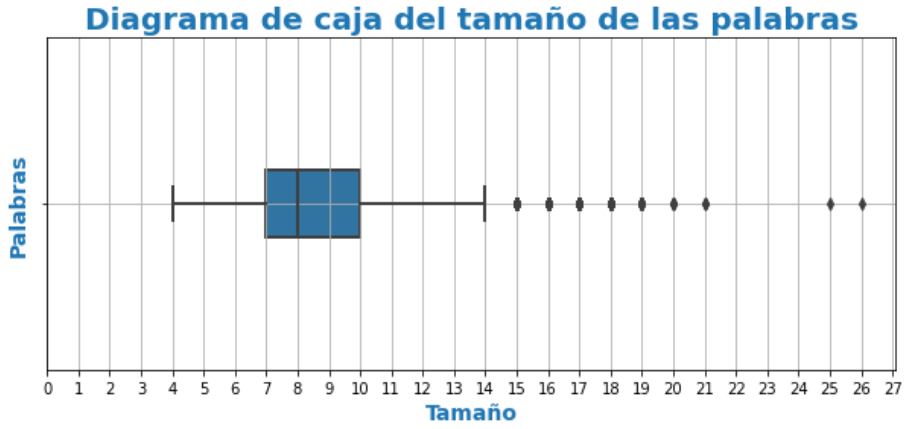
A continuación, se muestra una tabla con los diferentes cálculos, en estas se pueden evidenciar las diferentes maneras de dispersión respecto a cada letra, como se puede evidenciar la dispersión de cada letra respecto a la media en todos los datos es menor a 4

Descripción	Rango	Varianza	Desviación estándar
ñ	12	3.8704348703970117	1.9673420827088033
z	14	6.893261594849866	2.625502160511369
y	13	5.787877782969807	2.405800861037714
x	11	2.533272283272283	1.5916256731003944
w	11	4.860244030241785	2.204596114992899
v	14	7.2263348632609	2.6881843060439325
u	16	4.580320026829316	2.1401682239556115
t	21	8.561397438354993	2.9259865752178347
s	20	13.551833099884183	3.68128144806726
r	17	6.985912659044718	2.6430877130819397
q	13	4.3196162057772955	2.0783686404912136
p	17	4.064750681823465	2.016122685211261
o	24	10.682323091932844	3.268382335641417
n	23	9.946650194521176	3.1538310345548277
m	19	8.193274511597386	2.862389650553779
l	24	7.5334986832861395	2.7447219683031903
k	14	6.901151555499577	2.6270042930112574
j	13	5.030532428821051	2.242884845198489
i	20	7.607495340272334	2.7581688382461893
h	14	6.046034684787823	2.4588685781854673
g	15	5.637473841684713	2.3743365055704957
f	16	5.891309567942393	2.4272020039424804
e	25	10.147443496062063	3.1855052183385393
d	22	10.640752659546703	3.2620166553141177
c	16	9.0501956974695	3.0083543171424307
b	16	7.1024515309910035	2.6650425007851193
a	22	9.570495341245934	3.0936217191579733

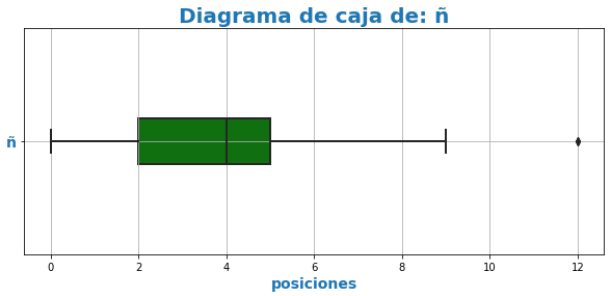
7. Exploración visual de los datos:

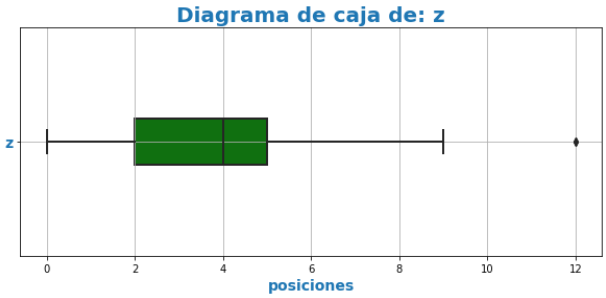
* Realizar gráficas que permitan comprender el comportamiento de la longitud de las palabras y la posición de los caracteres en cada una de ellas. Realizar boxplots, scatterplots y gráficas para variables categóricas de ser necesario.
* Realizar el análisis de las gráficas generadas.

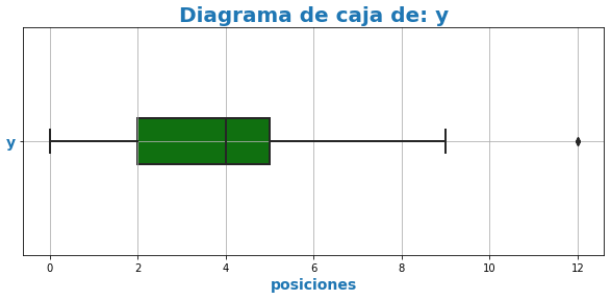
como anteriormente se mostró los valores de los percentiles ahora se muestra las graficas respecto a la longitud de cada palabra.

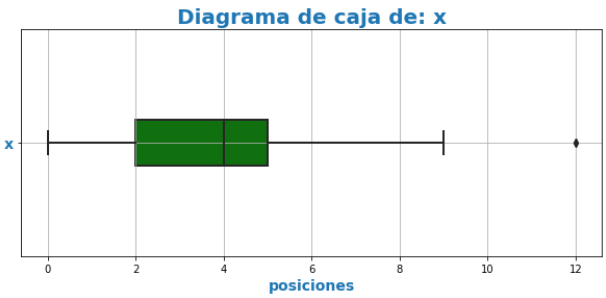


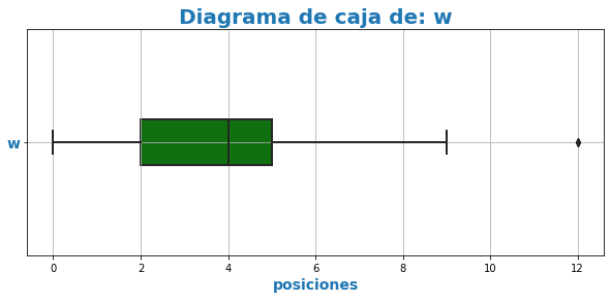
Ahora se muestra las graficas correspondientes a cada letra

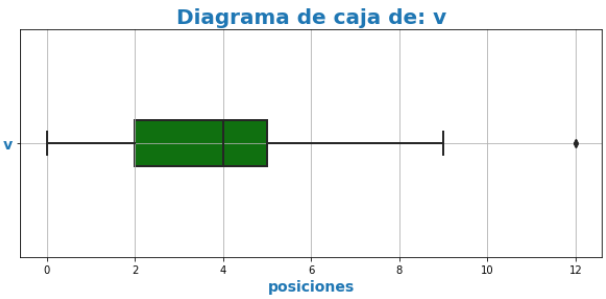


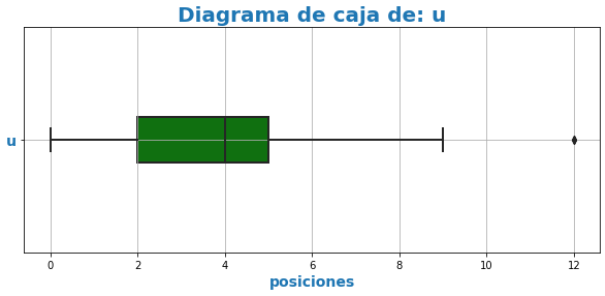


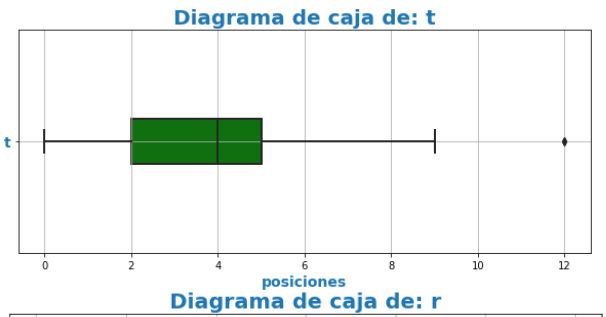


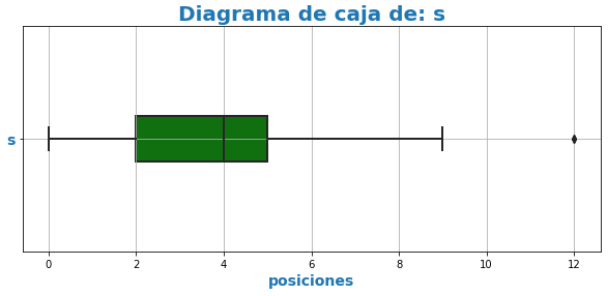


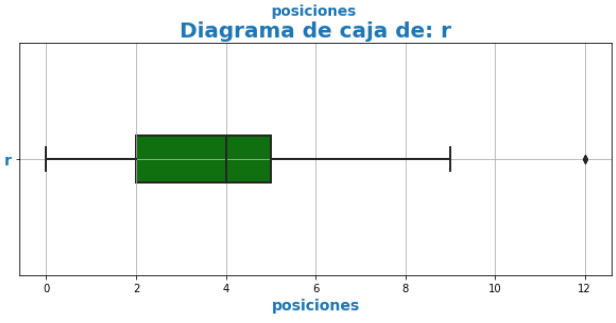












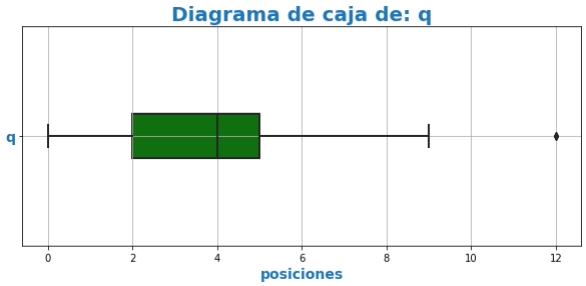


Diagrama de caja de: q
q
0
2
4
6
8
10
12
posiciones

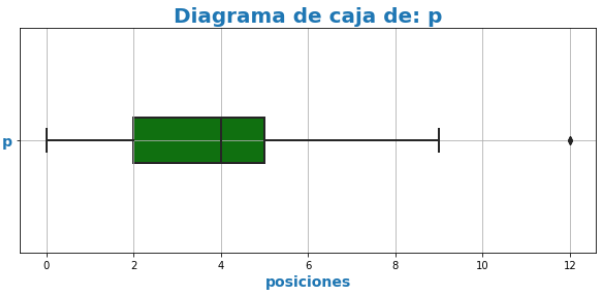


Diagrama de caja de: p
p
0
2
4
6
8
10
12
posiciones

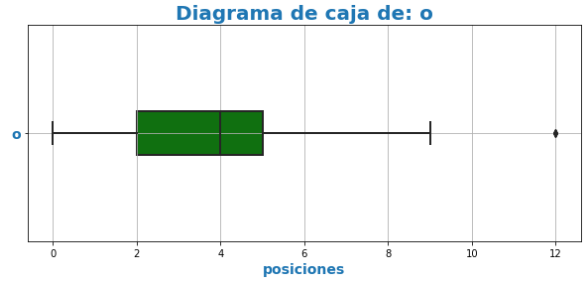


Diagrama de caja de: o
o
0
2
4
6
8
10
12
posiciones

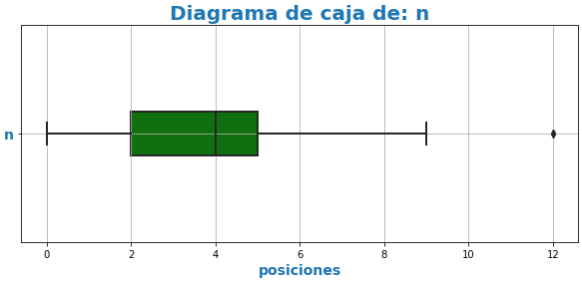


Diagrama de caja de: n
n
0
2
4
6
8
10
12
posiciones

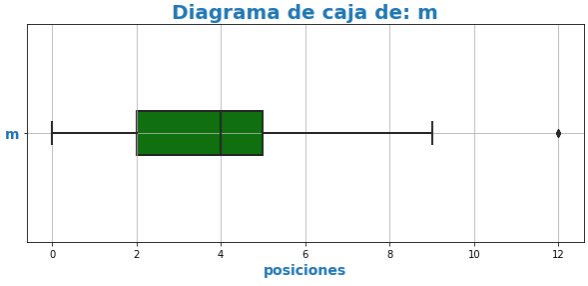


Diagrama de caja de: m
m
0
2
4
6
8
10
12
posiciones

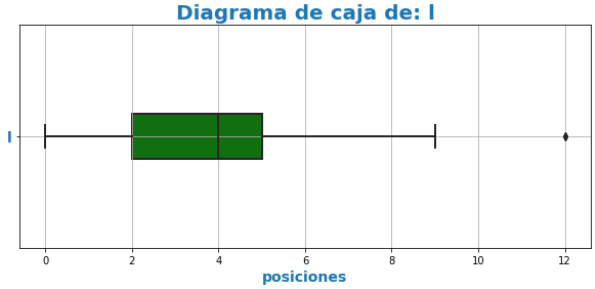


Diagrama de caja de: l
l
0
2
4
6
8
10
12
posiciones

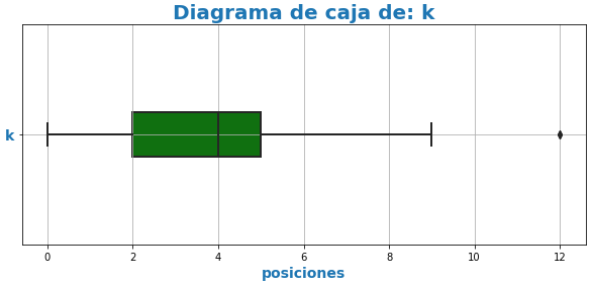


Diagrama de caja de: k
k
0
2
4
6
8
10
12
posiciones

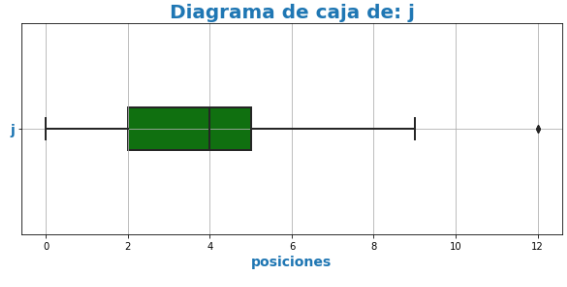


Diagrama de caja de: j
j
0
2
4
6
8
10
12
posiciones

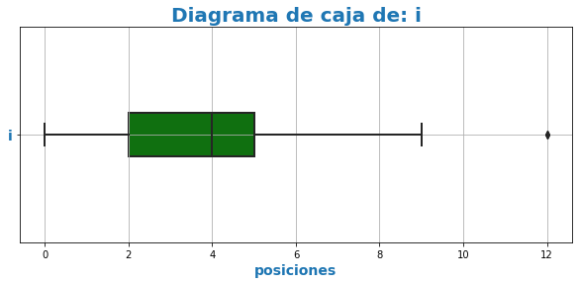


Diagrama de caja de: i
i
0
2
4
6
8
10
12
posiciones

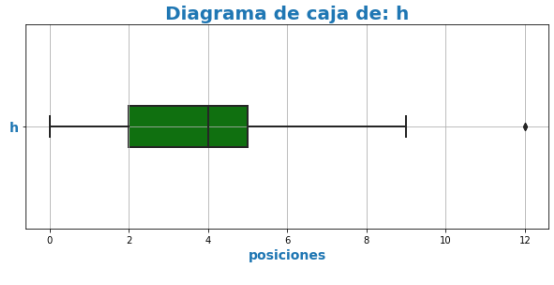


Diagrama de caja de: h
h
0
2
4
6
8
10
12
posiciones

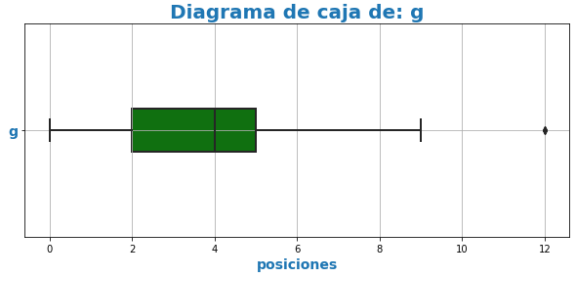


Diagrama de caja de: g
g
0
2
4
6
8
10
12
posiciones

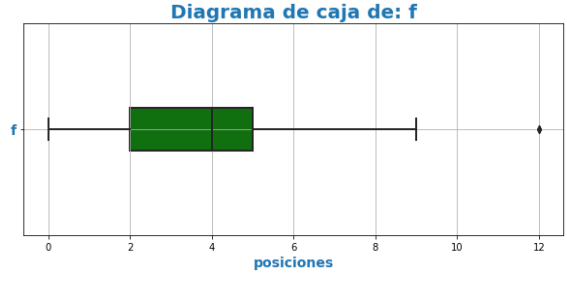
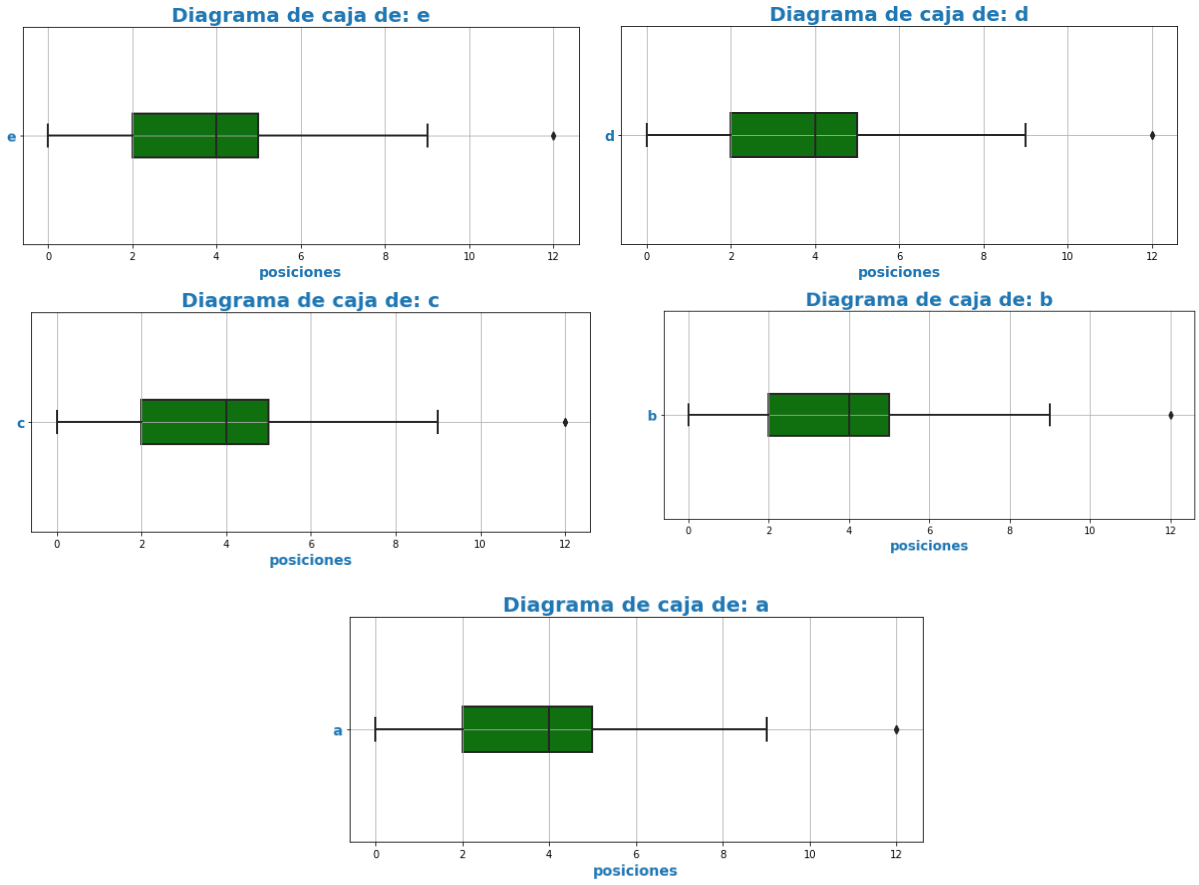
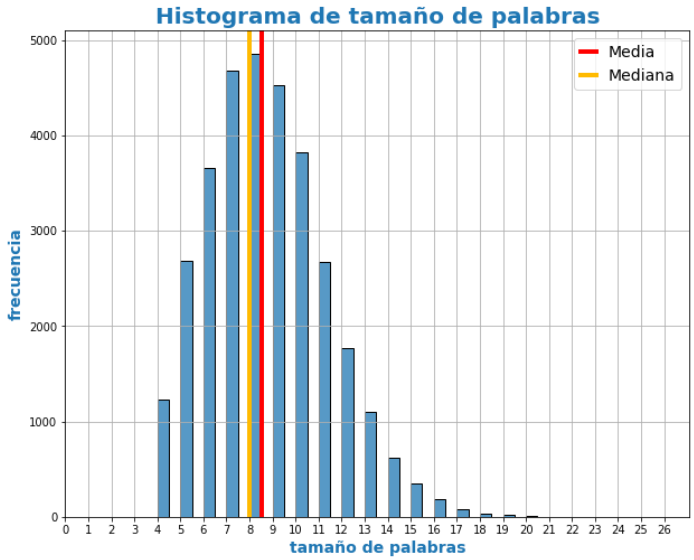


Diagrama de caja de: f
f
0
2
4
6
8
10
12
posiciones

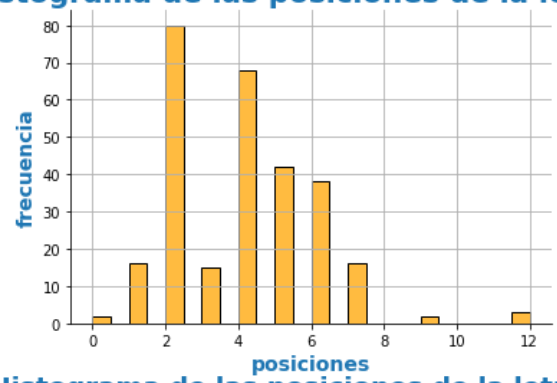


8. Distribuciones de datos:

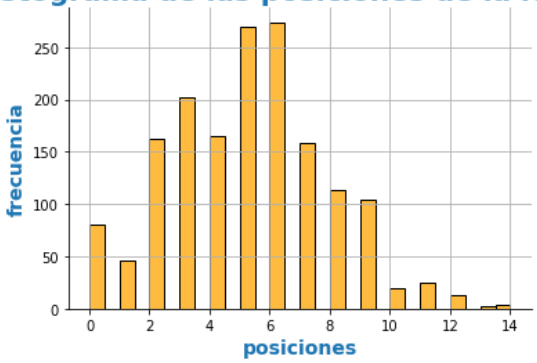
* Realizar gráficas de distribución para las longitudes de las palabras presentadas.
* Realizar gráficas de distribución para la posición de un carácter en específico en cada palabra.
* Realizar el análisis de las gráficas generadas.



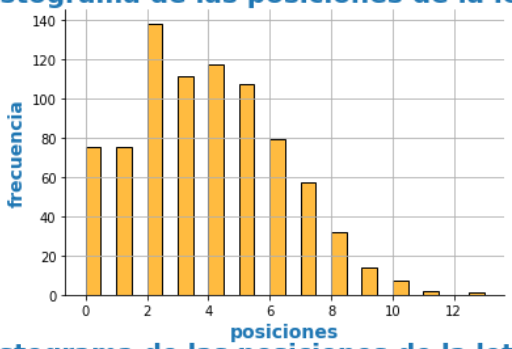
Histograma de las posiciones de la letra: ñ



Histograma de las posiciones de la letra: z



Histograma de las posiciones de la letra: y

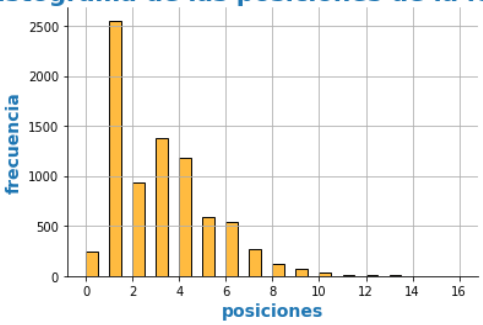


Histograma de las posiciones de la letra: x

Histograma de las posiciones de la letra: w

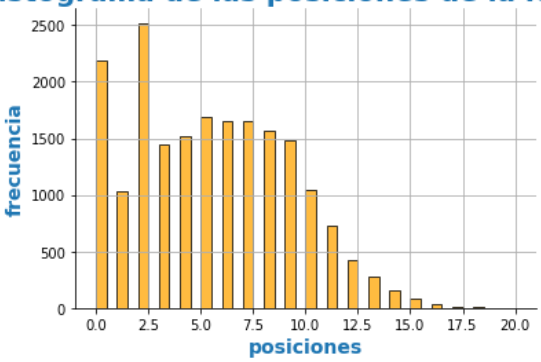
Histograma de las posiciones de la letra: v

Histograma de las posiciones de la letra: u

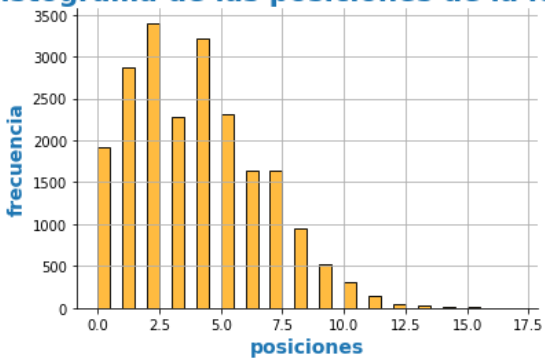


Histograma de las posiciones de la letra: t

Histograma de las posiciones de la letra: s

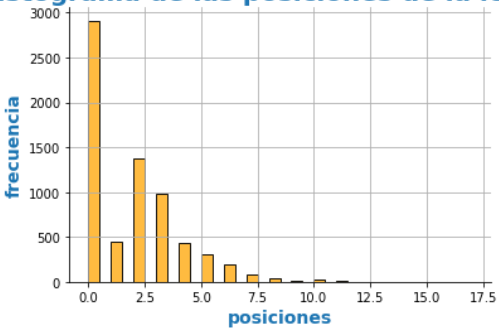


Histograma de las posiciones de la letra: r



Histograma de las posiciones de la letra: q

Histograma de las posiciones de la letra: p



Histograma de las posiciones de la letra: o

Histograma de las posiciones de la letra: n

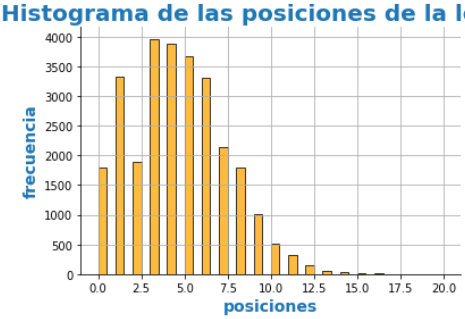
Histograma de las posiciones de la letra: m

Histograma de las posiciones de la letra: l

Histograma de las posiciones de la letra: k

Histograma de las posiciones de la letra: j

Histograma de las posiciones de la letra: i



Histograma de las posiciones de la letra: h

Histograma de las posiciones de la letra: g

Histograma de las posiciones de la letra: f

Histograma de las posiciones de la letra: e

Histograma de las posiciones de la letra: d

Histograma de las posiciones de la letra: c

Histograma de las posiciones de la letra: b

9. Introducción a las pruebas de hipótesis:

* Determinar si existe significancia estadística entre las posiciones en las que están las vocales y consonantes ubicadas en las palabras

Para hacer las pruebas de hipótesis se realizo la prueba de kolmogorov-smirnov para ello se hizo dos hipótesis la nula y la alternativa.
* H0 = Determinar si existe significancia estadística entre las posiciones en las que están las vocales y consonantes ubicadas en las palabras
* H1 = no existe significancia estadística entre las posiciones en las que están las vocales y consonantes ubicadas en las palabras

como los valores de p son menores al nivel de significancia 5% se rechaza la hipótesis nula. según lo obtenido se tiene la suficiente evidencia estadística para para decir que no existe significancia estadística entre las posiciones en las que están las vocales y consonantes ubicadas en las palabras.

10. Determinación de correlación entre variables:

* Determinar la correlación existente entre la longitud de las palabras y la cantidad de vocales contenidas en la misma.
* Determinar la correlación existente entre la longitud de las palabras y la cantidad de consonantes contenidas en la misma.

se concluyó que no se tiene una correlación estadística significativa entre las dos variables longitud de las palabras y letras del abecedario.

11. Verificación final de los datos:

* Demostrar en las salidas de información: la calidad de la información y que se estén cumpliendo las validaciones propuestas para los datos

La nueva base de datos cumple con la mayor parte de los requisitos mostrados a continuación:
1. No permitir palabras con más de una letra en mayúscula (Eliminarlas).
2. Reemplazar letras con tildes por sus homólogos sin ellas.
3. Reemplazar caracteres especiales por espacios, ej: ["\t", "\n", "\u200b"]
4. Permitir sólo letras del alfabeto español (Letras de la Aa a la Zz incluyendo la Ññ), reemplazar las demás por espacios en blanco.
5. Reemplazar cualquier símbolo (@, !, ?, etc...) por espacio en blanco.

6. Reemplazar stop words por espacio en blanco.
7. Sólo permitir palabras con más de 3 caracteres de longitud después de haber realizado la limpieza.
8. Finalmente todas las palabras deben estar en minúscula.
9. sin palabras repetidas

Con lo anterior, hay que tener en cuenta que frente a los requisitos planteados, puede la base de datos aun tenga algunas palabras por depurar, que están en ingles y comparten el mismo alfabeto que el español.