# LIMPIEZA Y VALIDACIÓN DE LOS DATOS

En esta práctica se elabora un caso práctico orientado a aprender a identificar los datos relevantes para un proyecto analítico, donde se aplicarán las técnicas específicas de tratamiento de datos (integración, transformación, limpieza y validación) para su posterior análisis. Se entregará un solo archivo con el enlace Github (<https://github.com>). Se utilizará la Wiki de Github para describir el proyecto.

El objetivo de esta actividad será el tratamiento de un dataset. Las diferentes tareas a realizar (y **justificar**) son las siguientes:

1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende responder?
2. Limpieza de los datos.
   1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?
   2. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos?

¿Cómo gestionarías cada uno de estos casos?

1. Análisis de los datos.
   1. Selección de los grupos de datos que se quieren analizar/comparar.
   2. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.
   3. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.
2. Representación de los resultados a partir de tablas y gráficas.
3. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
4. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos. También podéis trabajar en Python.

Hay que entregar un único fichero que contenga el enlace Github donde haya:

1. Una Wiki con el nombre del componente y una descripción de los ficheros.
2. Un documento Word, Open Office o PDF con las respuestas a las preguntas y con el nombre del componente.
3. Una carpeta con el código generado para analizar los datos.
4. El fichero CSV con los datos originales.
5. El fichero CSV con los datos finales analizados.

Este documento de entrega final de la Práctica 2 se tiene que entregar en el espacio de Entrega y Registro de AC del aula virtual antes de las **23:59** del día **8 de enero**.

## LIMPIEZA Y VALIDACIÓN DE LOS DATOS

1. Descripción del dataset ¿Por qué es importante y qué pregunta/problema pretende responder?
2. ¿Qué pregunta/problema pretende responder?

**¿Qué valen las mujeres?** [[1]](#footnote-1) **En los últimos quince años,** [**más de 900 hombres han asesinado en España**](http://www.tribunafeminista.org/2018/01/918-hombres-han-acabado-con-la-vida-de-sus-parejas-o-ex-parejas-en-15-anos/) **a sus parejas o ex parejas**. No estamos hablando de un problema residual, ni puntual.

Si nos fijamos en las estadísticas oficiales, cuya fuente es el Consejo General del Poder Judicial, y ofrecidas en el Portal Estadístico de la Delegación del Gobierno para la Violencia de Género (fecha de referencia: 30/09/2017), tenemos:[[2]](#footnote-2)

|  |  |
| --- | --- |
| **Año** | **Núm. de denuncias por violencia de género** |
| **Año 2009** | 135.539 |
| **Año 2010** | 134.105 |
| **Año 2011** | 134.002 |
| **Año 2012** | 128.477 |
| **Año 2013** | 124.893 |
| **Año 2014** | 126.742 |
| **Año 2015** | 129.193 |
| **Año 2016** | 143.535 |
| **Año 2017** | 125.769 |

|  |  |
| --- | --- |
| **Año** | **Núm. de órdenes de protección** |
| **Año 2009** | 41.081 |
| **Año 2010** | 37.908 |
| **Año 2011** | 35.813 |
| **Año 2012** | 34.537 |
| **Año 2013** | 32.831 |
| **Año 2014** | 33.167 |
| **Año 2015** | 36.292 |
| **Año 2016** | 37.958 |
| **Año 2017** | 29.455 |

|  |  |
| --- | --- |
| **Año** | **Número de víctimas mortales** |
| **Año 2003** | 71 |
| **Año 2004** | 72 |
| **Año 2005** | 57 |
| **Año 2006** | 69 |
| **Año 2007** | 71 |
| **Año 2008** | 76 |
| **Año 2009** | 56 |
| **Año 2010** | 73 |
| **Año 2011** | 62 |
| **Año 2012** | 52 |
| **Año 2013** | 54 |
| **Año 2014** | 55 |
| **Año 2015** | 60 |
| **Año 2016** | 44 |
| **Año 2017** | 48 |

Cuando se trata de abordar un problema tan grave y tan serio como los asesinatos machistas que se suceden año tras año, mes tras mes, semana tras semana, cabría preguntarse, qué valor le damos como sociedad a la vida de una mujer.

Salvo honrosas excepciones, los hombres siguen considerando el combate por la igualdad y contra la violencia de género como algo que atañe únicamente a las mujeres. Craso error. Se trata de una lucha en la que nos tenemos que embarcar todos para que tenga garantía de éxito.

El engranaje de la violencia machista es muy complejo y se compone de múltiples piezas. Su versatilidad es casi infinita porque el machismo es capaz de mutar y adaptarse al ritmo de los tiempos. Pero el daño que produce es siempre el mismo y tiene siempre como objetivo el sometimiento de la mitad de la sociedad, que son las mujeres.

Es necesario recorrer el camino hacia la plena igualdad, hacia la igualdad real entre hombres y mujeres, un camino que pase por la educación y que facilite el destierro del machismo de la sociedad.

**Necesitamos conocer, si existe en la sociedad una** actitud proactiva en contra de la violencia machista y a favor del feminismo (principio de igualdad de derechos de la mujer y el hombre y cuestionar la dominación y violencia de los varones sobre las mujeres) que pueda cambiar el estatus quo de las víctimas, o no existe esa actitud**.**

El feminismo es un fenómeno omnipresente en todo el mundo, y quizás donde más se debata, sea en Twitter y otras redes sociales. En la actualidad la influencia de las redes sociales es constante en la población de todo el mundo y quizás el hashtag #Feminismo haya creado tendencia social al ser capaz de expandir mensajes de toda índole.

Es por ello que, a través del hashtag **#Feminismo**, que es una forma de relacionar directamente los mensajes de los usuarios en Twitter, se podrán observar las opiniones y sentimientos que este tema provoca en la sociedad y podremos responder, tras analizar datos de twitter, qué y cómo se opina del feminismo.

1. Descripción del dataset

Las operaciones a realizar a través de **RStudio**, son las siguientes:

1. Acceder a la API pública de Twitter.

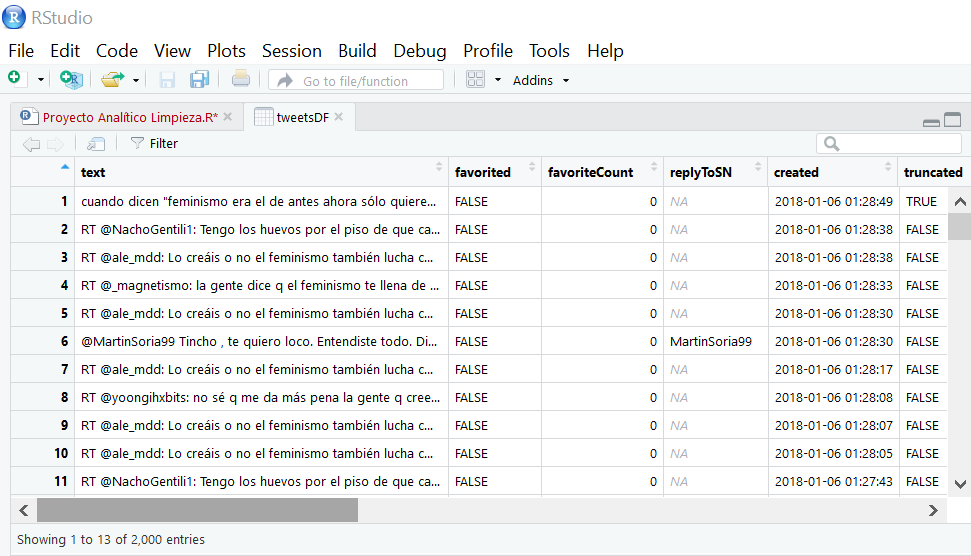


1. Descargar el conjunto de datos en un dataset (fuente de datos original).
2. Limpieza de los datos y crear el dataset para el análisis.
3. Análisis de los datos:
   1. Análisis de Sentimientos.
   2. Análisis estadísticos de los resultados
4. Representación de los resultados:
   1. Representación de las valoraciones por sectores
   2. Gráfico dinámico: 'highchart'
   3. Otros gráficos: histograma y boxplot.
5. Modelización y Predicción.
   1. Machine Learning: 'RTexTools'
   2. Recta de regresión y nube de puntos

El lenguaje de programación R permite a los ‘*data scientists*’ gestionar grandes volúmenes de datos estadísticos, pero necesita en ocasiones instalar y cargar determinadas librerías -paquetes adicionales que le dan una capacidad de gestión de datos-, por lo que en este caso será necesario cargar: library(twitteR), library(plyr), library(stringr), library(ggplot2), library(highcharter), library(RTextTools), library(e1071) y library(SparseM).

Para acceder a la API pública de Twitter, será necesario previamente gestionar los permisos correspondientes en su aplicación y obtener los parámetros de acceso: consumer\_key, consumer\_secret, access\_token y access\_secret.

Para obtener unos datos suficientemente representativos, se ha optado por descargar 2000 tweets, relacionados con el hashtag #Feminismo, con 16 variables por tweet. Posteriormente, los datos serán incorporados a un *data frame* y exportados a un archivo con formato *csv*, con el nombre: “*1-tweetsDF-ORIGINAL.csv*”.



*Muestra de 1 a 13 tweets, de los 2000 que contiene el archivo 1-tweetsDF-ORIGINAL.csv*

1. Limpieza de los datos.
   1. Selección de los datos de interés a analizar. ¿Cuáles son los campos más relevantes para responder al problema?

Es a través del hashtag #Feminismo, como los usuarios en Twitter relacionan directamente sus mensajes con cadena de caracteres formada por la palabra feminismo precedida por una almohadilla.

Dado que se necesita conocer, si en Twitter existe una actitud proactiva o no a favor del feminismo y la tendencia social que se expande tras el hashtag #Feminismo, es en el campo “text”, donde se podrán observar las opiniones y sentimientos que este tema provoca en la sociedad. Si los datos de interés a analizar están en el campo “text” de cada tweet, se puede descartar el resto de variables del tweet al no ser de interés para esta práctica.

Posteriormente con los datos del campo “text”, realizaremos las siguientes operaciones:

1. Limpieza del texto, mediante la aplicación de una función diseñada *ad hoc*.
2. Realizar un análisis de sentimientos, mediante la aplicación de la correspondiente función.
   1. ¿Los datos contienen ceros o elementos vacíos? ¿Y valores extremos? ¿Cómo gestionarías cada uno de estos casos?

Un **Tweet** es un mensaje que alcanzan hasta 280 caracteres: letras, números, signos y enlaces. Si en un tweet existen diferentes caracteres que no tienen significación sentimental alguna, en 2000 tweets la cifra es muy significativa, por ello eliminarlos supondría una laboriosa transformación sino se utilizase un procesado de texto.

En R existen varias funciones, si bien en este caso se ha utilizado *gsub()*, la cual reemplaza la aparición de una subcadena con otra subcadena dentro de un vector.

La función de limpieza del texto diseñada, que se puede consultar en el archivo del código generado en R, permitió eliminar en cada uno de los tweets: números, RT, espacios vacíos, caracteres propios de Twitter sin significación de sentimientos y los links. Como resultado de la limpieza de datos, el campo “text” estará en condiciones de serle aplicado el análisis de sentimientos en cada uno de los 2000 tweets.

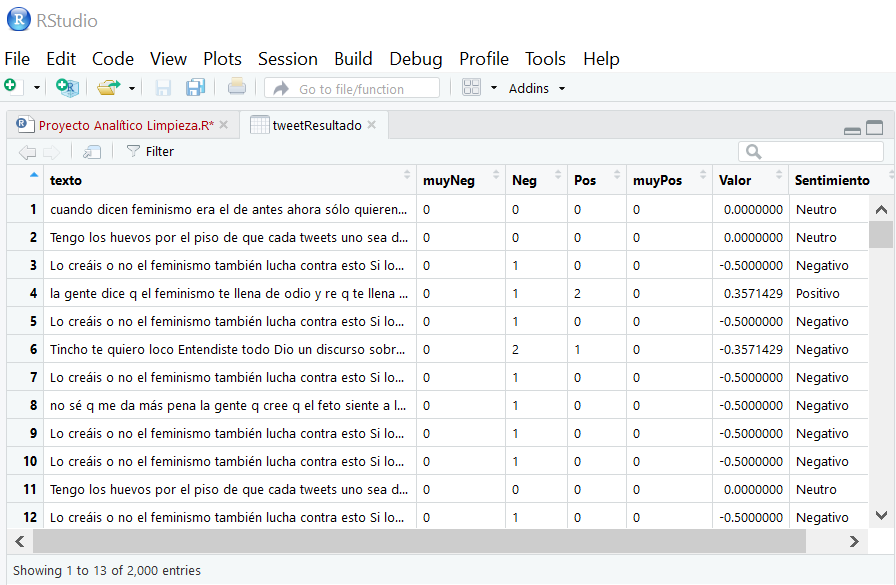
1. Análisis de los datos.
   1. Selección de los grupos de datos que se quieren analizar/comparar.

Puesto que el objetivo es analizar las palabras contenidas en el texto limpio de los tweets, será necesario disponer de un diccionario de palabras ponderadas según sentimientos relacionados con la palabra “Feminismo”, para comparar cada palabra de los tweets con las del diccionario y si coinciden asignarle el valor correspondiente.

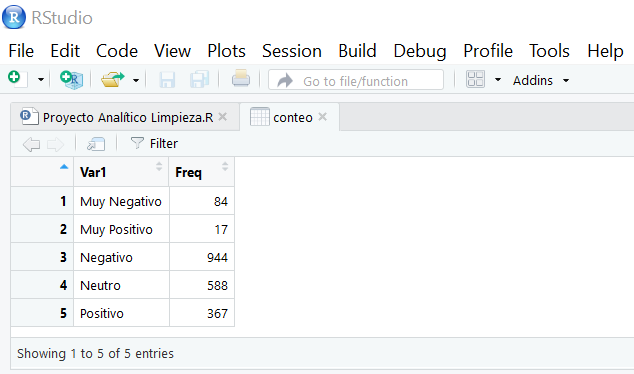
Señalar tres cuestiones: la primera, que el diccionario de sentimientos en origen corresponde a valoración de películas descargado de internet con muchos errores y palabras repetidas, por lo que se tuvo que limpiar previamente a su uso; la segunda, que las palabras incluidas en los tweets relacionados con el hashtag #Feminismo no resultaban significativas en los primeros análisis de sentimientos, siendo fundamentalmente el resultado final neutro a pesar de observar que no se correspondía con las opiniones del algunos tweets; y tercero, una vez analizadas las palabras de los tweets, 200 de ellas cargadas de significación con el hashtag #Feminismo fueron incorporadas al diccionario y otras 200 palabras incluidas en el diccionario fueron cambiadas del masculino al femenino (ejemplo: torturada por torturado, agredida por agredido, humillada por humillado, etc.) e incluso se en determinados casos se han mantenido las dos formas. El nombre del archivo del diccionario de sentimientos es: “*LISTA-Palabras-Sentimientos.txt*”

Una vez que se tiene el texto de los tweets limpio y el diccionario de sentimientos confeccionado, se crea una FUNCIÓN para clasificar palabras del texto de los tweet, en 4 categorías: muy Negativo, negativo, positivo y muy Positivo.

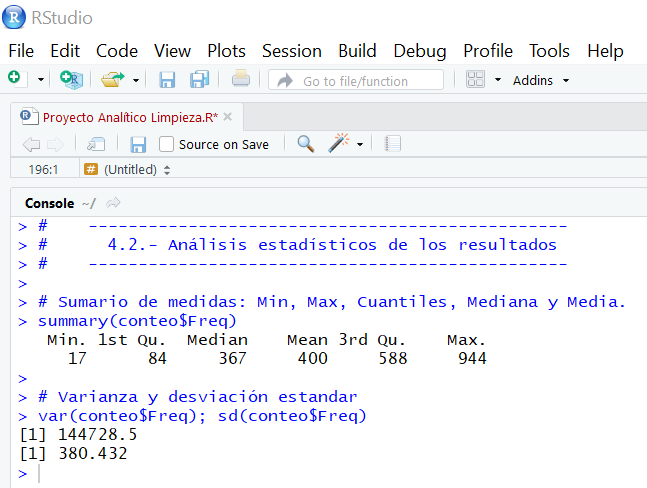
Seguidamente se creará una matriz “*scores\_final*”, la cual incluirá el campo “text” y los cuatro campos correspondientes a la clasificación de las palabras. Se calcula la puntuación de cada tweet y se añaden dos nuevas columnas, la del valor alcanzado por cada tweet y el del sentimiento asignado. El *data frame* se exporta a un archivo csv: “*2-tweet.Resultado.csv*".



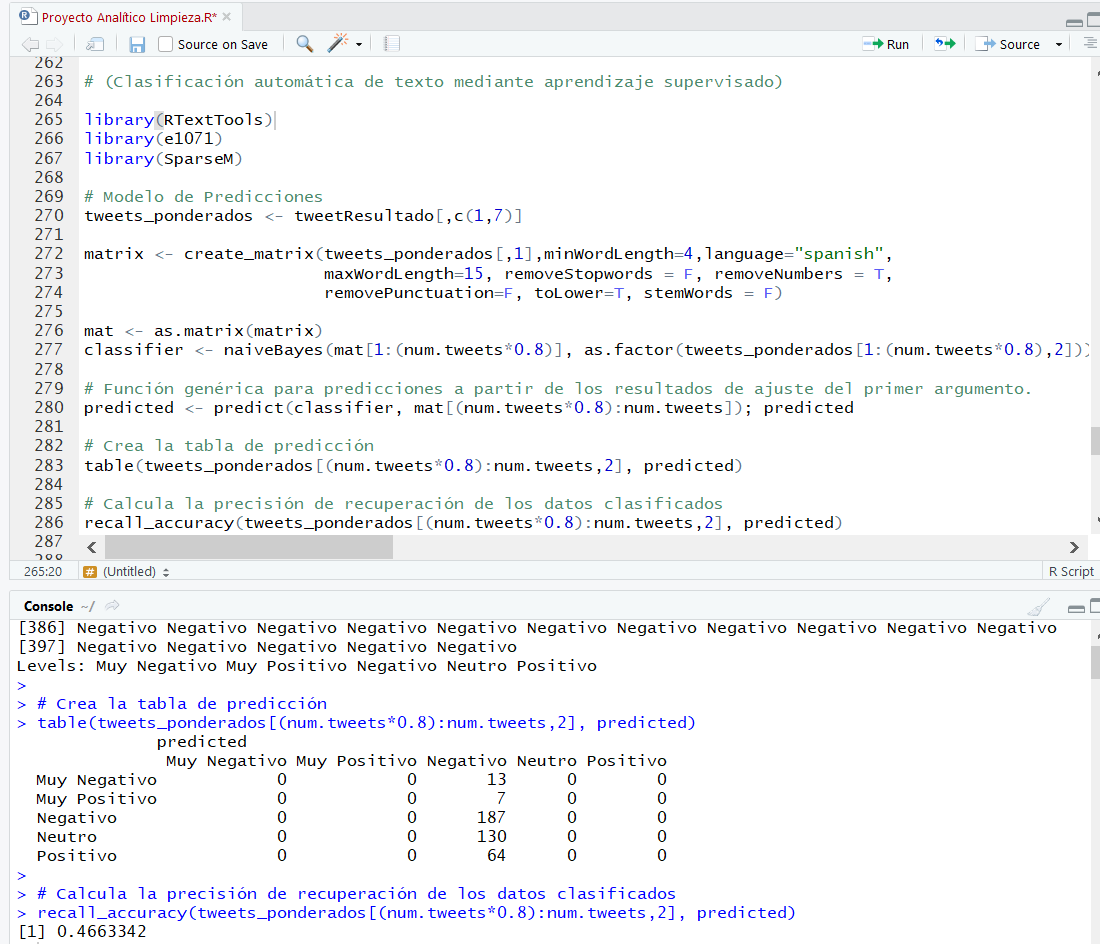
Finalmente se crea una tabla de sentimientos 5x2, correspondiendo a la primera columna sentimientos (cuatro categorías de sentimientos más el neutro -ausencia de sentimientos-) y a la segunda columnas la frecuencia observada para cada sentimiento. La tabla de sentimientos es exportada a un archivo csv: "*3-Tabla-Conteo-Sentimientos.csv*".



* 1. Comprobación de la normalidad y homogeneidad de la varianza. Si es necesario (y posible), aplicar transformaciones que normalicen los datos.

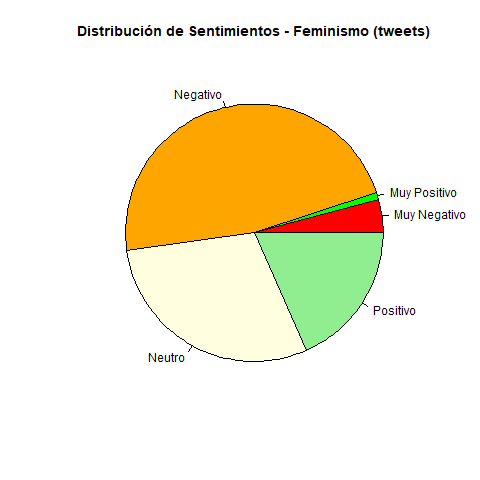


* 1. Aplicación de pruebas estadísticas (tantas como sea posible) para comparar los grupos de datos.

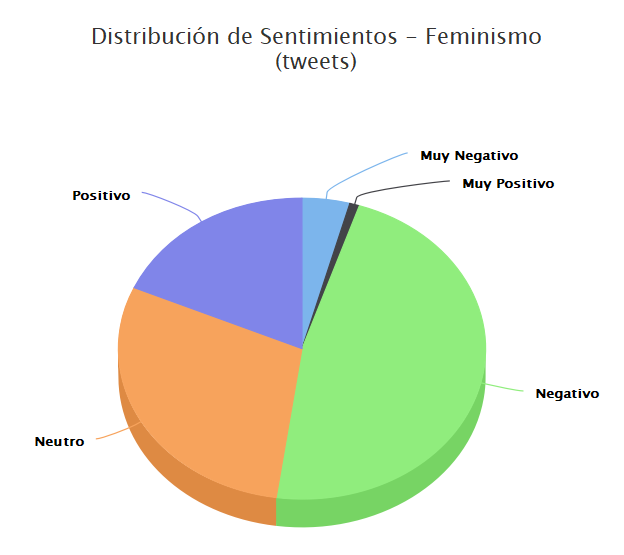


1. Representación de los resultados a partir de tablas y gráficas.

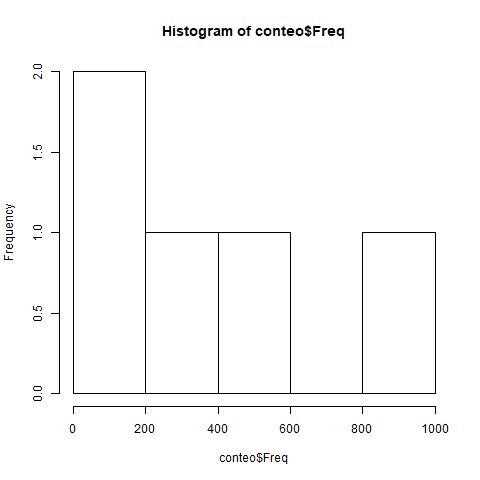
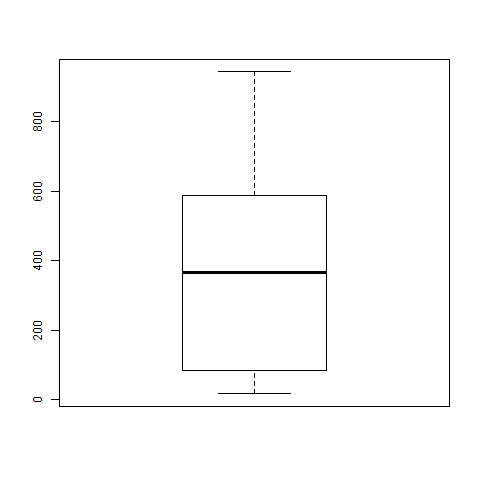
3.2-Rplot.Sentimientos.Feminismo.png



3-Rplot.Sentimientos.Feminismo



5.3.2-Boxplot.Feminismo.png 5.3.1-Hist.Feminismo.png



1. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?
2. Código: Hay que adjuntar el código, preferiblemente en R, con el que se ha realizado la limpieza, análisis y representación de los datos.

Ha sido subido a GitHub y se encuentra en el siguiente enlace:

<https://github.com/salgadogb/Tipologia_Practica_2>

El código está completamente comentado.

1. ¿Qué valen las mujeres? Lidia Guinart Moreno. Periodista y escritora. [*http://www.tribunafeminista.org/2018/01/que-valen-las-mujeres/*](http://www.tribunafeminista.org/2018/01/que-valen-las-mujeres/) [↑](#footnote-ref-1)
2. Portal Estadístico de la Delegación del Gobierno para la Violencia de Género. [*http://estadisticasviolenciagenero.msssi.gob.es/*](http://estadisticasviolenciagenero.msssi.gob.es/) [↑](#footnote-ref-2)