Tipología y Ciclo de Vida de los Datos: PRA2 - Limpieza y validación de los datos

Autores: Joel Bustos - Iván Ruiz

Junio 2020

Table of Contents

[Introducción 1](#_Toc40807691)

[Presentación. Práctica 2. 2](#_Toc40807692)

[1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende resolver? 2](#_Toc40807693)

[2. Integración y selección de los datos de interés a analizar. 4](#_Toc40807694)

[• 2.1 Resumen de tratamientos previos 11](#_Toc40807695)

[• 2.2 Carga del nuevo archivo tras el preprocesado 13](#_Toc40807696)

[3. Limpieza de los datos. 17](#_Toc40807697)

[• 3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos? 17](#_Toc40807698)

[• 3.2 Identificación y tratamiento de valores extremos 17](#_Toc40807699)

[4. Análisis de los datos 17](#_Toc40807700)

[• 4.1 Selección de los grupos de datos que se quieren analizar/compararar (planificación de los análisis a aplicar) 18](#_Toc40807701)

[• 4.2 Comprobación de la normalidad y homogeneidad de la varianza 18](#_Toc40807702)

[• 4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos 18](#_Toc40807703)

[5. Representación de los resultados a partir de tablas y gráficas 18](#_Toc40807704)

[6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema? 18](#_Toc40807705)

[Contribuciones 18](#_Toc40807706)

[Bibliografía 18](#_Toc40807707)

# Introducción

## Presentación. Práctica 2.

A lo largo de esta segunda práctica, vamos a tratar de profundizar en el análisis que los autores este trabajo planteamos en la primera parte de la asignatura. Así, hacemos referencia al repositorio que creamos y a la documentación generada durante el primer trabajo:

<https://github.com/iruiper/Cyberattacks-History>

En concreto, a pesar de las dificultades que plantea el set de datos sobre el que trabajaremos (y tal como expondremos a través de las distintas secciones), sí que nos gustaría tratar de cerrar las inquietudes y motivación que nos llevó en primera instancia a trabajar sobre la problemática de los ataques de ciberseguridad. En este sentido, nos gustaría iniciar esta segunda exposición rescatando la motivación que planteábamos en el proyecto de obtención de los datos:

*Los equipos de seguridad han necesitado incorporar, cada vez más, perfiles técnicos en el área de la ciberseguridad. Estos equipos técnicos, normalmente con un conocimiento muy específico, en ocasiones no disponen de demasiadas herramientas que les permitan ser proactivos y anticiparse a las nuevas tendencias y técnicas de ciberataque. De esta forma, acaban adaptando un comportamiento reactivo, realizando tareas de mantenimiento y de respuesta ante incidentes.*

*Nos planteábamos, como contexto para la presente práctica, recopilar datos históricos de ciberataques con el objetivo de crear un modelo predictivo que sirviese de soporte al equipo de seguridad de una empresa. Idealmente, estudiando lo que está ocurriendo en relación a delitos cibernéticos, los equipos internos de las distintas entidades, podrían tratar de prepararse mejor contra aquellos riesgos a los que los modelos estadísticos les pudieran sugerir que se encuentran más expuestos.*

Con ese punto de partida, como decimos, a continuación vamos a tratar de plantear un problema concreto que creemos que podría analizarse a través de los datos que hemos conseguido rascar del sitio web <https://www.hackmageddon.com/>.

Para abordar este segundo proyecto, veremos que los datos brutos extraídos durante la primera práctica conllevan retos y problemas iniciales, tales como la falta de agregación, limpieza y homogenización, o incluso la falta de variables relevantes para el análisis. Dado que, como decimos, hay carencias de calidad de datos de base, a pesar de que la propuesta de práctica se centra en tareas de limpieza y acondicionado mediante R, utilizaremos una combinación de Python y R por los motivos que iremos exponiendo en los distintos puntos del tratamiento de los datos.

# 1. Descripción del dataset. ¿Por qué es importante y qué pregunta/problema pretende resolver?

Tal y como explicábamos en la introducción anterior, disponemos de un set de datos con información histórica y clasificada sobre ataques cibernéticos con gran impacto que han producido desde el año 2017 hasta la actualidad. Con esta información, cabe preguntarse hasta qué punto podemos anticiparnos a los problemas que sobrevendrán a una entidad, o si estos datos pueden ayudar en asuntos financieros y presupuestarios para responder a preguntas tales como:

* ¿Hasta qué punto mi compañía se encuentra más expuesta a ciberataques por el sector en el que se encuentra?
* ¿Debería dotar presupuesto y medios adicionales en algún periodo particular del año?
* ¿Debería analizar la documentación cualitativa y los detalles técnicos de algún tipo de ataque particular que afecte a mi compañía?
* ¿Hasta qué punto el encontrarme en un mundo globalizado me expone a ataques internacionales?
* ¿Existen atacantes bien conocidos, con pautas concretas, a los que me encuentre particularmente expuesto?

Todas las preguntas anteriores están relacionadas con buena parte de la información que sí tenemos ya recabada a partir del set de datos en bruto que obtuvimos en durante la realización de la primera práctica:

* Disponemos de datos sobre las entidades o sectores afectados por los ciberataques que aparecen documentados en el dataset.
* Disponemos de la fecha en la que se han producido los distintos ataques.
* Sabemos si los ataques son de amplia cobertura, dado que sabemos si el ataque se produjo sobre un país concreto, o sobre la agrupación de varios de ellos.
* Sabemos qué ataques se encuentran bien documentados en cuanto a su origen (atacante identificado), versus ataques de difícil trazabilidad (atacante desconocido).
* Conocemos la tipología de ataque a que corresponden los distintos incidentes reportados.

En concreto, para que los resultados y los contrastes que llevemos a cabo sean lo más concretos y realistas posibles, vamos a centrarnos en el caso de que **formemos parte del equipo de seguridad de un organismo público**, por lo que nuestros análisis cuantitativos y cualitativos tratarán poner de relieve las diferencias entre nuestro sector y los demás. Esta distición también puede tener como derivada interesante, averiguar el nivel de gasto e inversión acometido por entidades de otros sectores en materia de ciberseguridad, y a partir de nuestra evaluación del riesgo específico, estudiar si puede ser necesaria la aplicación de nuevas partidas presupuestarias para la defensa contra estas amenazas.

Como hemos visto en los materiales didácticos de Subirats, Pérez y Calvo [1], existen muy diferentes retos a la hora de integrar y asegurar la calidad de los datos que necesitamos para dar repuesta a las inquietudes de cualquier analítica de datos. También, según se desprende de dicho material, y de los ejemplos basados en sets de datos estructurados, gran parte de los problemas vienen por los datos numéricos que se utilizan en los estudios. Sin embargo, la principal dificultad que ofrece el set de datos que hemos elegido es precisamente que lo que tenemos no son estadísticas numéricas o datos agregados, sino detalle de incidentes individuales. En consecuencia, gran parte del reto al que nos enfrentamos, y gran parte de los problemas de limpieza que vamos a desarrollar a lo largo de la práctica irán dirigidos a “crear” datos numéricos y estadísticos que permitan resolver las cuestiones sobre las que queremos obtener respuestas.

# 2. Integración y selección de los datos de interés a analizar.

El primer reto que deberemos resolver es precisamente la integración e incluso la creación de las variables numéricas que podrán ayudarnos a resolver las preguntas que nos hemos marcado como objetivo en la sección anterior. Si recordamos, como resultado de nuestra primera práctica, fuimos capaces de recopilar el detalle de incidentes reportados y analizados en el sitio web <https://www.hackmageddon.com/>.

No pretendemos repetir toda la exposición sobre el proceso de extracción, pero es importante explicar y entender que los datos recabados sufren gran heterogeneidad por los distintos motivos

* Los incidentes más actuales (año 2019 en adelante) se reportan mediante informes quincenales, con estructura específica.
* Los incidentes del año 2018 se encuentran agregados en forma tabular.
* Los incidentes del año 2017 se encuentran agregados en forma tabular, pero con formato distinto al año 2018.

En consecuencia, el primer problema con el que nos encontramos (antes incluso de poder generar variables numéricas que cuantifiquen el tipo de ataques), es que tenemos varior archivos csv para distintos periodos temporales, y que el formato de los campos en cada uno de ellos no tiene por qué ser necesariamente igual. A continuación vamos a recordar la estructura de campos del set de datos generado, y vamos a comprobarlo sobre una concatenación sin procesos de tratamiento o limpieza adicionales (llamaremos a este frame *attacks\_Raw*).

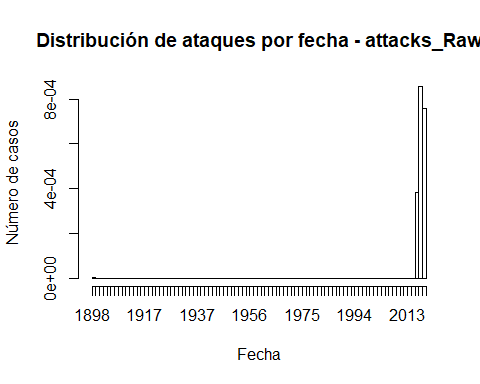
# Almacenamos el set de datos bruto en el frame "attacks\_Raw" para un análisis preliminar de algunos de los campos que utilizaremos  
attacks\_Raw <- read.csv2(file='DatosAtaques\_2017\_2020\_RAW.csv',stringsAsFactors = TRUE)  
  
# Mostramos la estructura del archivo recién cargado  
str(attacks\_Raw)

## 'data.frame': 4468 obs. of 11 variables:  
## $ ID : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ Date : Factor w/ 1054 levels "01/01/2017","01/01/2018",..: 1 1 1 1 71 104 104 104 104 138 ...  
## $ Author : Factor w/ 564 levels ""," LulzSecITA",..: 17 7 166 7 38 7 289 7 7 186 ...  
## $ Target : Factor w/ 3518 levels ""," City of Del Rio",..: 2547 2954 1070 2743 1385 1417 1231 969 2080 2528 ...  
## $ Description : Factor w/ 4453 levels ""," Malaysia's Computer Emergency Response Team (MyCERT) reveal the details of a campaign carried out by APT40, ta"| \_\_truncated\_\_,..: 3591 3443 1072 4046 92 1412 1522 1040 3895 1885 ...  
## $ Attack : Factor w/ 327 levels "\"view as\" vulnerability",..: 3 280 220 177 91 177 97 177 9 263 ...  
## $ Target.Class: Factor w/ 25 levels "C Manufacturing",..: 13 2 13 16 13 13 7 16 15 23 ...  
## $ Attack.Class: Factor w/ 13 levels " CC",">1","CC",..: 4 6 3 3 10 4 3 3 3 4 ...  
## $ Country : Factor w/ 158 levels "",">1","AE","AF",..: 47 129 139 139 47 59 16 139 139 2 ...  
## $ Link : Factor w/ 1144 levels "","http://abcnews.go.com/Politics/fbi-probing-attempted-hack-trump-organization-officials/story?id=47652150",..: 268 174 68 714 1042 44 1032 724 815 69 ...  
## $ Tags : Factor w/ 927 levels "","#OpIsrael, #OpUSA, Anonymous",..: 352 831 213 754 381 556 456 261 582 245 ...

Observamos, ya sobre esta carga inicial, que tenemos varios problemas que resolver:

* Existen campos con mucha información cualitativa, pero que no resultan relevantes para el estudio de los problemas concretos que hemos marcado como objetivo. En consecuencia, podemos comenzar con procesos de **reducción de la dimensionalidad**, a través del descarte de aquellos atributos que no vamos a necesitar:
  + ID. Identificador único dentro de cada informe.
  + Target. Nombre concreto de la entidad atacada, pero nuestro estudio trata de agregar tipologías o sectores.
  + Description. Información de detalle de cada incidente, pero nuestro estudio debe agregar necesariamente tipos de ataques, por lo que el detalle individual no servirá a efectos estadísticos.
  + Attack. De manera análoga a “Description”, es un campo con información detallada sobre el tipo de ataque, pero utilizaremos tipologías agregadas para el análisis.
  + Link. Ofrece un enlace la URL en la que se puede estudiar el detalle que ofrece una noticia sobre el incidente reportado, pero no será de utilidad en el proyectos de explotación de datos que estamos planteando.
  + Tags. Contiene hashtags que resumen categorías, pero de nuevo ofrece información demasiado granular para el objetivo que nos hemos marcado.
* El campo fecha (**Date**) tiene un nivel de granularidad excesivo, ya que no resultará muy dificil encontrar varios ataques reportados en el mismo día. El nivel de granularidad que vamos a marcar para cada observación del set de datos será el número de ataques reportados en un mes concreto, por lo que será necesario separar este campo en dos: “Anyo” y “Mes”.
* Otro problema adicional de la fecha es que podría haberse registrado mal en origen. Podemos hacer un par de breves análisis sobre este campo, que nos confirmarán la necesidad de analizar y limpiar dicho campo:

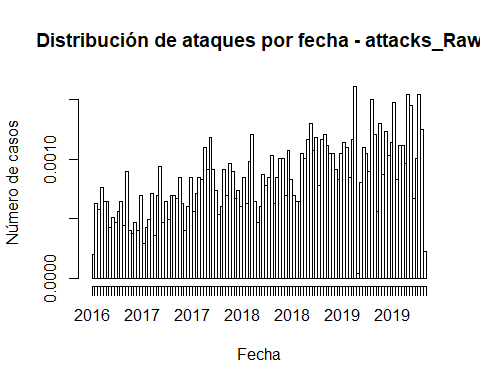
# Extracción y conversión del campo que registra la fecha del incidente  
fechas <- as.Date(attacks\_Raw$Date, format="%d/%m/%Y")  
  
# Análisis gráfico de la distribución de incidentes por fecha  
hist(fechas, breaks=100, main="Distribución de ataques por fecha - attacks\_Raw", xlab = "Fecha", ylab="Número de casos")



# Se observa que necesariamente hay valores erróneos (el típico registro de valor nulo y/o conversión a 01/01/1900). Veámoslo en modo tabla.  
table(format(fechas,"%Y"))

##   
## 1900 2017 2018 2019 2020   
## 4 951 1338 1693 482

# Analicemos de nuevo la distribución de casos por fecha si eliminamos los casos de 1900  
hist(fechas[format(fechas,"%Y")!="1900"], breaks=100, main="Distribución de ataques por fecha - attacks\_Raw", xlab = "Fecha", ylab="Número de casos")



* En relación al tipo de ataque (**Attack.Class**), lo que vamos a tratar de analizar es si hay alguna tipología de ataque concreto que afecte en mayor medida a nuestra entidad, o poder estudiar qué tipo de relación puede existir en cuestión de tendencias, correlación entre distintos tipos de ataques, etc. Para ello, nuestra propuesta consiste en la **creación de variables numéricas** que recojan el número de ataques por cada tipo, así como el número total de ataques por mes y año. Sin embargo, tenemos que analizar primero la calidad de datos de esta variable:

# Analizaremos la calidad de los datos de la variable Tipo de Ataque  
  
tipoAtaque <- attacks\_Raw$Attack.Class  
table(tipoAtaque)

## tipoAtaque  
## CC >1 CC CE   
## 1 1 1469 232   
## CW CW? Cyber Crime Cyber Espionage   
## 50 1 1094 172   
## Cyber Warfare H Hacktivism N/A   
## 33 58 38 5   
## Not Found   
## 1314

* Sobre la tabla anterior, observamos que necesitaremos tendremos que llevar a cabo algunas tareas:
* Homogenizar los valores. Vemos que encontramos tanto valores “Cyber Crime” como “CC”, o “Cyber Warfare” como “CW”. Deberemos acordar un criterio, que será el de las iniciales, y crearemos una estructura tipo *crosstab* [2] con la distribución de ataques por tipo, teniendo en cuenta que tendremos una observación por cada año, mes y tipo de entidad (comentaremos a continuación las particularidades que contempla el tipo de entidad).
* Vemos que hay valores atípicos como “>1” o “N/A”, que agruparemos en “Otros”.
* Observamos la existencia de **valores ausentes**. En los procesos de obtención de datos, es posible que se produzcan errores por distintos motivos, o que incluso en el propio informe de la web no se haya registrado ese dato. Sea por el motivo que sea, hay que tomar una decisión sobre qué hacer con dichos valores, como se nos indica en [1]. Una de las posibilidades pasaría por completar los datos manualmente, como indica Osborne [3], o volver a corregir las tareas del proceso de obtención (podría servir para mejora la calidad de los procesos de obtención de datos, corrigiendo o contemplando nuevas particularidadesd desde la fase de *scraping*).
* La manera de distinguir estos casos va a ser incluir estos ataques en la sección de “Otros”, pero utilizaremos también una variable dicotómica “ProblemasQC”, y que podremos incluso utilizar en nuestros análisis estadísticos para tratar de averiguar si las características de otras observaciones podrían ayudar a tomar una decisión sobre el tipo de ataque a que podría corresponder el valor que no pudo obtenerse.
* En relación al tipo de entidad (**Target.Class**), analizaremos con la misma metodología la necesidad de tener en cuenta algún tipo de consideración particular.

# Analizaremos la calidad de los datos de la variable Entidad Atacada  
  
tipoEntidad <- attacks\_Raw$Target.Class  
table(tipoEntidad)

## tipoEntidad  
## C Manufacturing   
## 82   
## D Electricity gas steam and air conditioning supply   
## 46   
## E Water supply, sewerage waste management, and remediation activities   
## 7   
## G Wholesale and retail trade   
## 86   
## H Transportation and storage   
## 38   
## I Accommodation and food service activities   
## 71   
## J Information and communication   
## 201   
## K Financial and insurance activities   
## 193   
## L Real estate activities   
## 4   
## M Professional scientific and technical activities   
## 68   
## N Administrative and support service activities   
## 33   
## Not Found   
## 1314   
## O Public administration and defence, compulsory social security   
## 263   
## O Public administration, defence, compulsory social security   
## 167   
## P Education   
## 201   
## Q Human health and social work activities   
## 281   
## R Arts entertainment and recreation   
## 119   
## S Other service activities   
## 57   
## U Activities of extraterritorial organizations and bodies   
## 25   
## V Fintech   
## 88   
## X Individual   
## 672   
## Y Multiple Industries   
## 352   
## Y Multiple targets   
## 78   
## Y Multiple Targets   
## 9   
## Z Unknown   
## 13

* Sobre la tabla anterior, observamos que necesitaremos tendremos que llevar a cabo algunas tareas:
* Existe un identificador único (Primera letra), que nos indica algún valor que no tiene exactamente el mismo texto. De hecho, el caso de la entidad que queremos analizar (organismos públicos) está afectada por este error de **calidad de los datos**: “O Public administration and defence, compulsory social security” y “O Public administration, defence, compulsory social security”. Será necesario, por tanto, tener en cuenta como identificador único la letra de la categoría, y utilizar un descriptivo único.
* Como decíamos anteriormente, vamos a considerar un caso como el número de ataques sufridos por un tipo de organismo por mes, por lo que necesitaremos que la estructura *crosstab* que utilicemos considere como claves de reparto del número de ataques estos tres atributos: **Anyo**, **Mes** y **TipoEntidad**.
* Volvemos a tener el problema de **valores perdidos**, que contienen el valor “Not Found”. De nuevo, utilizaremos la dicotómica “ProblemasQC” para conservar el número de casos, y tendremos que crear una entidad *dummy* que denominaremos “\_ Errores” (sustituimos la letra mayúscula de valor único por el caracter guión bajo).
* Al igual que comentábamos el caso de necesidad de homogenización en la entidad “O”, apreciamos que también es necesario hacer lo mismo en “Y”.
* En relación al autor (**Author**), dado el alto número de posibles valores (564 niveles, como veíamos anteriormente), lo que nos parece más interesante aquí es distinguir, en cada observación, distinguir cuántos de los ataques son de delincuentes u organizaciones bien identificadas, y cuántos corresponden a autores anónimos o desconocidos. Esta distinción podría ayudarnos a identificar si estamos más expuestos a ataques de redes conocidas, y por lo tanto dedicar recursos a analizar sus sistemas y tipos de ataques, o si necesitamos mecanismos de defensa mucho más heterogénos porque el número de atacantes diversos sea alto. Veamos cómo podríamos crear dos variables cuantitativas que recojan el **número de ataques de atacantes conocidos**, y **número de ataques de atacantes desconocidos**.

# Analizaremos la calidad de los datos de la variable Author  
head(sort(table(attacks\_Raw$Author), decreasing = T), n = 20)

##   
## ? Anonymous   
## 3612 25   
## Russia? APT28 AKA Fancy Bear   
## 15 14   
## China? TA505   
## 12 12   
## OurMine The Dark Overlord   
## 11 11   
## Turla   
## 10 9   
## Lazarus Group Magecart   
## 9 9   
## APT28 FIN7   
## 8 8   
## APT10 Gnosticplayers   
## 7 7   
## AnonPlus Hidden Cobra   
## 6 6   
## LulzSec ITA and AntiSec ITA MuddyWater   
## 6 6

* Sobre la tabla anterior, observamos que necesitaremos tener en cuenta las siguientes consideraciones:
* En la estructura *crosstab*, la variable “**CasosAutorDesconocido**” contendrá los casos para los casos “?” y " " de la tabla anterior.
* El resto de casos sí que tienen un valor asignado de autor conocido, por lo que el resto de autores sumarán en la variable “**CasosAutorConocido**”.
* Por último, en relación al país afectado (**Country**), observamos que el número de casos documentados es muy alto (158 niveles), por lo que hemos decidido que para el estudio que queremos hacer, será suficiente segmentar el número de casos de ámbito local, versus número de casos con impacto internacional. Para ello, vamos a crear una variable dicotómica “**ImpactoGlobal**” que nos diga si para el tipo de entidad, mes y año, se ha producido algún caso con impacto internacional (valor binario). Veamos a partir de los posibles valores de la variable país, cómo considerar ese valor verdadero o falso:

# Analizaremos la calidad de los datos de la variable Country  
head(sort(table(attacks\_Raw$Country), decreasing = T), n = 20)

##   
## US >1 UK CA IT AU IN RU KR N/A DE JP FR BR   
## 1594 1460 160 90 82 63 63 58 51 50 48 47 44 38 35   
## CN UA IL HK NL   
## 34 29 26 22 21

* A partir de la tabla anterior, consideraremos que ha habido casos de impacto internacional en cada registro que construyamos, si encontramos casos con país “>1” en la combinación Entidad/Año/Mes que compone cada observación.

# 2.1 Resumen de tratamientos previos

Como indicábamos en la sección introductoria, en el proyecto de limpieza de datos que hemos planteado, puede que uno de los principales retos con nuestro set de datos, sea precisamente todo lo relativo a **integración**, **selección**, **reducción** y **conversión** de datos, porque partimos de una situación en la que ni siquiera tenemos una información totalmente tabulada y con las variables cuantitativas que necesitaríamos para el estudio.

En consecuencia, y dado que resulta necesario dar un paso atrás, y combinar las fuentes “*raw*” que creamos en la primera práctica, hemos considerado más eficiente y eficaz llevar a cabo los preprocesamientos anteriores mediante Python y la fuente de datos “DatosAtaques\_2017\_2020\_RAW.csv” que consolida todos los archivos recreados mediante *scraping* en la primera práctica.

Tal y como subimos al GitHub, complementa al código dinámico de R que incluye el presente informe, el código “**XXX.py**” **[por desarrollar]**, que viene a tratar de resolver todos los problemas que hemos expuesto y analizado anteriormente. Como resultado, hemos generado el fichero que nos acompañará en el resto del estudio, y que denominamos “**EstadisticasAtaques2017\_2020\_Input.csv**” **pendiente de crear**

Como hemos comentado anteriormente, las variables que componen cada observación del dataset que utilizaremos para limpiar y analizar son las siguientes *modificar con lo que finalmente componga el set de datos que genera Joel con Python*:

* **Anyo**. Año en el que se se documenta el número de ataques que recoge la observación.
* **Mes**. Mes en el que se se documenta el número de ataques que recoge la observación.
* **Entidad**. Tipo de entidad para la que se documenta el número de ataques, fabricada a partir de la variable original “Target.Class”.
* **ImpactoGlobal**. Variable dicotómica, verdadero o falso, indica si en ese tipo de entidad, en el mes y año correspondiente, se han producido ataques de impacto internacional. Atributo creado a partir de la variable original “Country”.
* **CasosAutorConocido**. Número de ataques con autor conocido e identificable que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de la variable original “Author”.
* **CasosAutorDesconocido**. Número de ataques con autor desonocido e identificable que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de la variable original “Author”.
* **AtaquesCC**. Número de ataques de tipo “Cyber Crime” que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **AtaquesCE**. Número de ataques de tipo “Cyber Espionage” que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **AtaquesCW**. Número de ataques de tipo “Cyber Warfare” que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **AtaquesH**. Número de ataques de tipo “Hacktivism” que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **AtaquesOtros**. Número de ataques de tipo “Otros” (valores “N/A”, “>1” o “Not Found”) que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **AtaquesTotal**. Número de ataques de tipo “Cyber Crime” que se han producido que se han producido para ese tipo de entidad, en el mes y año correspondiente. Variable creada a partir de conteo de casos, a través de la variable original “Attack.Class”.
* **ProblemasQC**. Identificador que señala si la observación concreta ha tenido problemas de calidad identificados en la etapa de generación de la información. Nos indica que los valores cuantificados pueden ser imprecisos, estar clasificados en categorías genéricas, o en entidades no identificadas.

# 2.2 Carga del nuevo archivo tras el preprocesado

Como hemos referido anteriormente, las extracciones en bruto de que disponíamos, no ofrecían un formato ni estructura adecuados para los objetivos del estudio, por lo que ha sido necesario dar un paso atrás y llevar a cabo tareas preliminares de limpieza y acondicionamiento desde el propio origen de los datos. Como hemos explicado, para ello ha resultado más eficiente y eficaz realizar algunas de estas tareas como mejora/modificación de los *scripts* desarrollados en la primera práctica. Y a tal efecto incluimos el código “**XXX.py**” **[por desarrollar]** en GitHub.

A continuación cargaremos el nuevo fichero de *input*, y explicaremos brevemente cómo vamos a avanzar en las siguientes secciones con dichos datos.

# Almacenamos el nuevo set de datos en el frame "attacks\_Input" para una validación adicional, y para explicar algunos de los tratamientos que llevaremos a cabo a lo largo del análisis  
attacks\_Input <- read.csv2(file='processed\_data.csv',stringsAsFactors = TRUE)  
attacks\_Input$Year <- as.factor(attacks\_Input$Year)  
attacks\_Input$Mes <- as.factor(attacks\_Input$Mes)  
  
# Creamos una nueva variable con el número total de ataques por observación  
attacks\_Input$NumeroAtaques <- attacks\_Input$Code\_attack\_class\_.1+attacks\_Input$Code\_attack\_class\_CC+attacks\_Input$Code\_attack\_class\_CE+attacks\_Input$Code\_attack\_class\_CW+attacks\_Input$Code\_attack\_class\_H+attacks\_Input$Code\_attack\_class\_UK  
  
# Mostramos la estructura del archivo recién cargado  
str(attacks\_Input)

## 'data.frame': 1748 obs. of 16 variables:  
## $ Year : Factor w/ 4 levels "2017","2018",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Mes : Factor w/ 12 levels "1","2","3","4",..: 1 1 1 1 1 1 1 1 1 1 ...  
## $ Continent : Factor w/ 7 levels "África","América",..: 2 2 2 2 2 2 2 2 2 2 ...  
## $ Country\_name : Factor w/ 112 levels "Afganistán","Alemania",..: 15 15 18 33 33 33 33 33 33 33 ...  
## $ Code\_target\_class : Factor w/ 21 levels "C","D","E","G",..: 7 21 13 1 6 7 10 11 12 13 ...  
## $ Desc\_target\_class : Factor w/ 21 levels "Accommodation and food service activities",..: 11 19 5 12 1 11 15 3 16 5 ...  
## $ ProblemasQC : logi FALSE FALSE FALSE FALSE FALSE FALSE ...  
## $ Author\_processed\_Conocido : int 1 0 0 0 0 1 0 0 1 0 ...  
## $ Author\_processed\_Desconocido: int 0 1 1 3 3 5 1 1 4 10 ...  
## $ Code\_attack\_class\_.1 : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Code\_attack\_class\_CC : int 1 1 1 3 3 6 1 1 5 10 ...  
## $ Code\_attack\_class\_CE : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Code\_attack\_class\_CW : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Code\_attack\_class\_H : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ Code\_attack\_class\_UK : int 0 0 0 0 0 0 0 0 0 0 ...  
## $ NumeroAtaques : int 1 1 1 3 3 6 1 1 5 10 ...

Podemos hacer una comprobación sobra una de las variables más relevantes para nuestros análisis numéricos: el número de ataques totales que estamos analizando, y segmentados en las distintas categorías de ataques que nos indica el atributo original “Attack.Class”.

# Análisis de integridad en la carga: comprobación de la completitud del dataset y la consistencia entre variables numéricas.  
  
colSums(attacks\_Input[7:15])

## ProblemasQC Author\_processed\_Conocido   
## 228 785   
## Author\_processed\_Desconocido Code\_attack\_class\_.1   
## 3899 2   
## Code\_attack\_class\_CC Code\_attack\_class\_CE   
## 2786 468   
## Code\_attack\_class\_CW Code\_attack\_class\_H   
## 104 1206   
## Code\_attack\_class\_UK   
## 118

cat("Suma del número de ataques por categoría: ",sum(colSums(attacks\_Input[8:15])[3:8]),"\n")

## Suma del número de ataques por categoría: 4684

cat("Suma del número de ataques de acuerdo a si se conoce el atacante: ",sum(colSums(attacks\_Input[8:15])[1:2]),"\n")

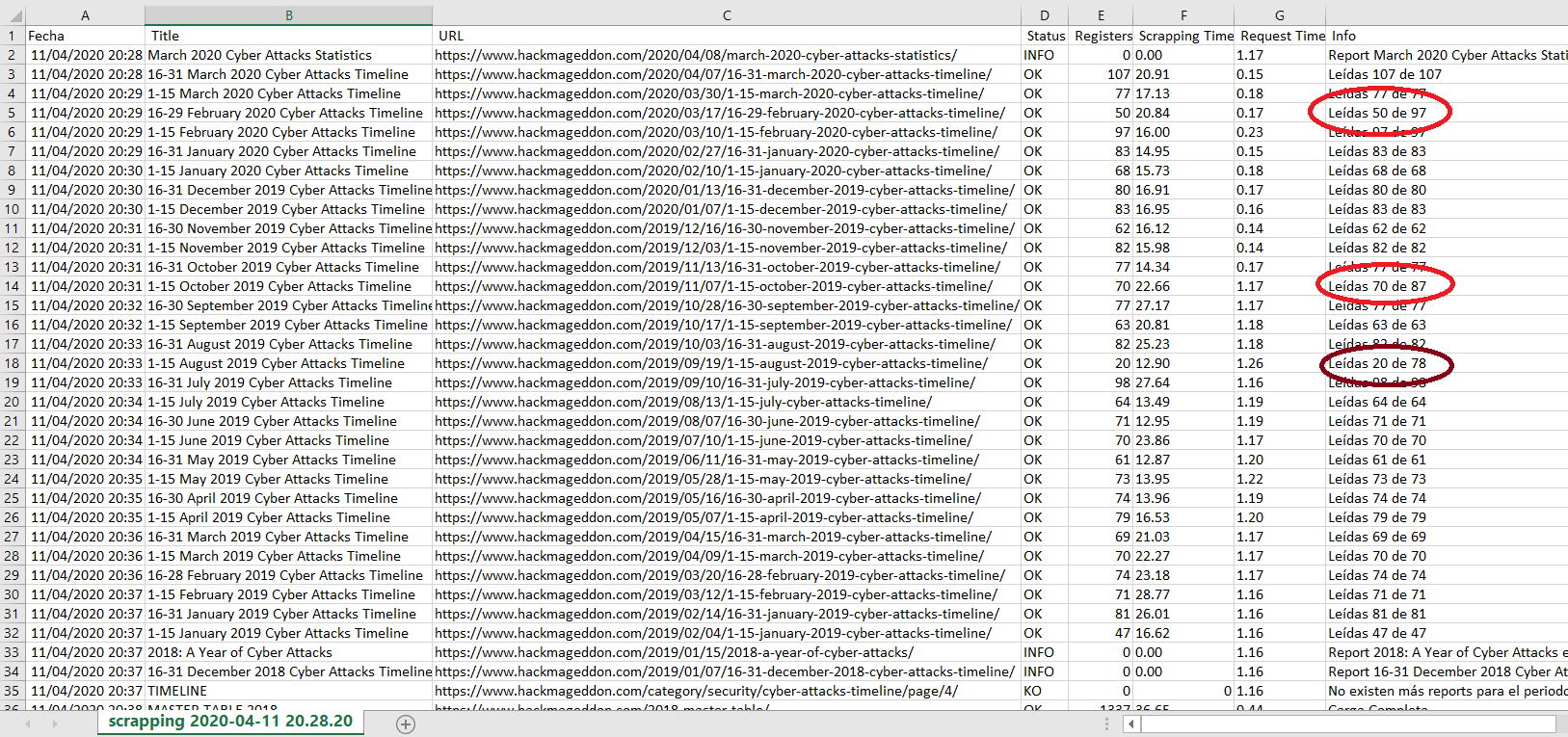
## Suma del número de ataques de acuerdo a si se conoce el atacante: 4684

cat("Valor acumulado en la variable NumeroAtaques: ",sum(attacks\_Input$NumeroAtaques))

## Valor acumulado en la variable NumeroAtaques: 4684

Con lo anterior, ya podemos considerar que tenemos un primer set de datos completo, y limpio de errores de procesamiento, aunque deberemos acabar de analizar si existen otros problemas de valores extremos a lo largo de la siguiente sección. No obstante, gracias al trabajo de preprocesamiento, hemos conseguido resolver algunos problemas importantes:

* Hemos llevado a cabo tareas de acondicionamiento para poder disponer de cuantificaciones basadas en Año/Mes/País/Entidad.
* Como veíamos al principio del análisis, teníamos muchísimas observaciones que constituían valores perdidos, por fallos en el proceso de extracción. Gracias a los **mecanismos de control de errores y de control de calidad** que hemos desarrollado en la primera práctica y que hemos complementado en la presente, conseguimos de alguna manera la mejor solución al caso de datos perdidos: recuperarlos manualmente. En esta ocasión el proceso no es del todo manual, sino que consiste en volver a ejecutar el *scraper* en los informes que había fallado, y de este modo completamos los datos que aparecían como “Not Found”.
* Mostramos a continuación un ejemplo del archivo de *log* generado en el proceso de rascado de datos, que ha permitido identificar aquellos casos en los que no se estaban leyendo todos los incidentes disponibles en el informe (resaltado sobre la captura algún ejemplo).



* Como resultado, no solo corregimos los valores ausentes, sino que hemos conseguido aumentar el número de casos a considerar en el análisis, desde los 4.468 de que disponíamos en el archivo inicial “DatosAtaques\_2017\_2020\_RAW.csv”, hasta los 4.684 que hemos conseguido recuperar tras el rascado adicional de los informes con error.
* Con el preprocesamiento hemos conseguido también reducir el número de errores en la identificación del tipo de ataque (inicialmente teníamos 1.314 casos “Not found”, y ahora tenemos únicamente 118 ataques documentados como “Unknown”.)
* Para el caso de las entidades, sigue habiendo ciertos problemas que no se han podido resolver en el preprocesamiento, pero al menos hemos podido identificarlos claramente, y podremos utilizarlos en nuestros análisis, siempre que tengamos la cautela de que son casos para los que se desconoce el tipo de entidad (inicialmente teníamos 1.314 casos “Not found”, y ahora tenemos 295 casos, pero que siguen correspondiendo a un número de ataques, 1.397, muy similar al que se apreciaba inicialmente como observamos en la consulta siguiente.)

# Analizaremos la calidad de los datos de la variable Entidad Atacada tras las tareas de preprocesamiento  
  
tipoEntidad <- attacks\_Input$Desc\_target\_class  
table(tipoEntidad)

## tipoEntidad  
## Accommodation and food service activities   
## 46   
## Activities of extraterritorial organizations and bodies   
## 25   
## Administrative and support service activities   
## 29   
## Arts entertainment and recreation   
## 81   
## Education   
## 71   
## Electricity gas steam and air conditioning supply   
## 43   
## Financial and insurance activities   
## 138   
## Fintech   
## 78   
## Human health and social work activities   
## 77   
## Individual   
## 169   
## Information and communication   
## 118   
## Manufacturing   
## 59   
## Multiple Industries   
## 104   
## Other service activities   
## 52   
## Professional scientific and technical activities   
## 51   
## Public administration and defence, compulsory social security   
## 227   
## Real estate activities   
## 6   
## Transportation and storage   
## 34   
## Unknown   
## 276   
## Water supply, sewerage waste management, and remediation activities   
## 6   
## Wholesale and retail trade   
## 58

colSums(attacks\_Input[attacks\_Input$Desc\_target\_class=="Unknown",8:16])

## Author\_processed\_Conocido Author\_processed\_Desconocido   
## 216 1181   
## Code\_attack\_class\_.1 Code\_attack\_class\_CC   
## 1 97   
## Code\_attack\_class\_CE Code\_attack\_class\_CW   
## 59 18   
## Code\_attack\_class\_H Code\_attack\_class\_UK   
## 1104 118   
## NumeroAtaques   
## 1397

# 3. Limpieza de los datos.

# 3.1 ¿Los datos contienen ceros o elementos vacíos? ¿Cómo gestionarías cada uno de estos casos?

# 3.2 Identificación y tratamiento de valores extremos

# 4. Análisis de los datos

# 4.1 Selección de los grupos de datos que se quieren analizar/compararar (planificación de los análisis a aplicar)

# 4.2 Comprobación de la normalidad y homogeneidad de la varianza

# 4.3 Aplicación de pruebas estadísticas para comparar los grupos de datos

# 5. Representación de los resultados a partir de tablas y gráficas

# 6. Resolución del problema. A partir de los resultados obtenidos, ¿cuáles son las conclusiones? ¿Los resultados permiten responder al problema?

## Contribuciones

|  |  |
| --- | --- |
| Investigación Previa Redacción de las respuestas Desarrollo código | JBP – IRP JBP – IRP JBP – IRP |

## Bibliografía

[1] Subirats, Laia - Pérez, Diego O. - Calvo, Mireia (2019). “Introducción a la limpieza y análisis de los datos”, Universidad Oberta de Catalunya

[2] Bock, Tim (2019). “What is a Crosstab”, Display R Blog. [en línea] [Última consulta: 15/May/2020] <https://www.displayr.com/what-is-a-crosstab/>

[3] Osborne, Jason W. (2013). “Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data.” Thousand Oaks, CA. Sage Publications.